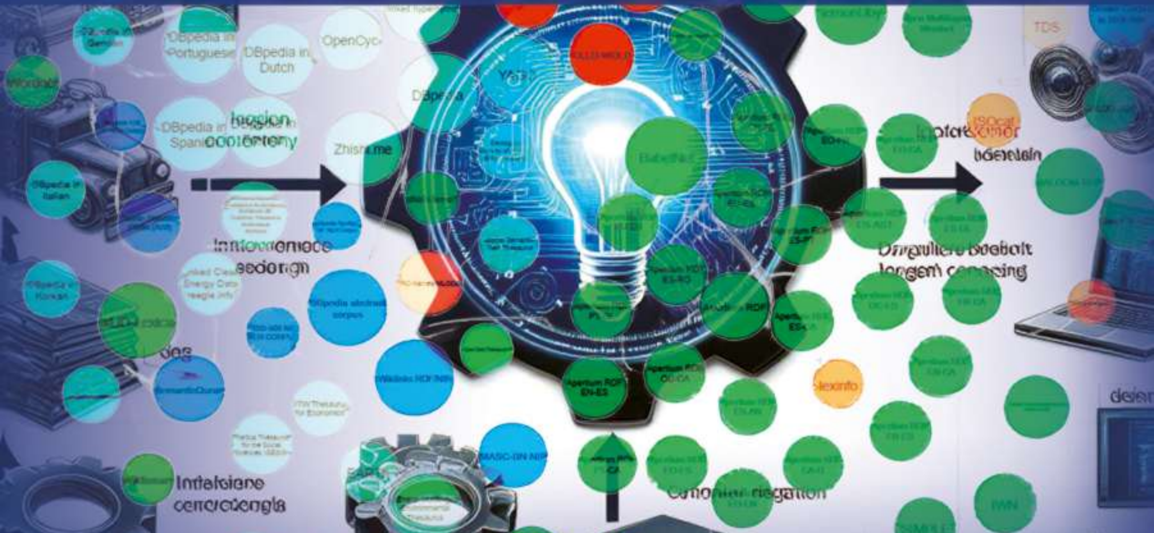
NATURAL LANGUAGE PROCESSING FOR

SOFTWARE ENGINEERING



Edited By

Rajesh Kumar Chakrawarti, Ranjana Sikarwar,
Sanjaya Kumar Sarangi, Samson Arun Raj Albert Raj, Shweta Gupta,
Krishnan Sakthidasan Sankaran and Romil Rawat

Scrivener Publishing

WILEY

Natural Language Processing for Software Engineering

Scrivener Publishing
100 Cummings Center, Suite 541J
Beverly, MA 01915-6106

Publishers at Scrivener
Martin Scrivener (martin@scrivenerpublishing.com)
Phillip Carmical (pcarmical@scrivenerpublishing.com)

Natural Language Processing for Software Engineering

Edited by

Rajesh Kumar Chakrawarti

Ranjana Sikarwar

Sanjaya Kumar Sarangi

Samson Arun Raj Albert Raj

Shweta Gupta

Krishnan Sakthidasan Sankaran

and

Romil Rawat

Scrivener
Publishing

WILEY

This edition first published 2025 by John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA
and Scrivener Publishing LLC, 100 Cummings Center, Suite 541J, Beverly, MA 01915, USA

For more information about Scrivener publications please visit www.scrivenerpublishing.com.

Wiley Global Headquarters
111 River Street, Hoboken, NJ 07030, USA

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Limit of Liability/Disclaimer of Warranty
While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials, or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read.

Library of Congress Cataloging-in-Publication Data

ISBN 9781394272433

Front cover images supplied by Adobe Firefly
Cover design by Russell Richardson

Set in size of 11pt and Minion Pro by Manila Typesetting Company, Makati, Philippines

Printed in the USA

10 9 8 7 6 5 4 3 2 1

Contents

Preface

The book's goal is to discuss the most current trends in applying natural language processing (NLP) approaches. It makes the case that these areas will continue to develop and merit contributions.

The book focusses on software development that is based on visual modelling, is object-orientated, and is one of the most significant development paradigms today. To reduce issues throughout the documentation process, there are still a few considerations to make. To assist developers in their documentation tasks, a few aids have been developed. To aid with the documentation process, a variety of related tools (such as assistants) may be made using natural language processing (NLP). The book is focused on software development and operation using data mining, informatics, big data analytics, artificial intelligence (AI), machine learning (ML), digital image processing, the Internet of Things (IoT), cloud computing, computer vision, cyber security, Industry 4.0, and health informatics domains.

1

Machine Learning and Artificial Intelligence for Detecting Cyber Security Threats in IoT Environmment

Ravindra Bhardwaj[1]*, Sreenivasulu Gogula[2], Bidisha Bhabani[3], K. Kanagalakshmi[4], Aparajita Mukherjee[5] and D. Vetrithangam[6]

[1]Deparment of Physics and Computer Science, Dayalbagh Educational Institute (Deemed to be University), Agra, Uttar Pradesh, India
[2]Department of CSE (Data Science), Vardhaman College of Engineering, Shamshabad, Hyderabad, India
[3]Department of Computer Science and Engineering, University of Engineering and Management (UEM), New Town, West Bengal, India
[4]Department of Computer Applications, SRM Institute of Science and Technology (Deemed to be University), Trichy, India
[5]Department of Computer Science and Engineering, Institute of Engineering and Management, University of Engineering and Management (UEM), New Town, Kolkata, West Bengal, India
[6]Department of Computer Science & Engineering University, Institute of Engineering, Chandigarh University, Mohali, Punjab, India

Abstract

The Internet of Things (IoT) refers to the increasing connectivity of many human-made entities, such as healthcare systems, smart homes, and smart grids, through the internet. Currently, a vast amount of material and expertise has been widely spread. These networks give rise to several security threats and privacy concerns. Intrusions refer to malevolent and unlawful actions that cause harm to the network. IoT networks are susceptible to a diverse range of security issues due to their widespread presence. Cyber attacks on the IoT architecture can lead to the loss of information or data, as well as the sluggishness of IoT devices. For the past twenty years, an Intrusion Detection System has been utilized to ensure the security of

**Corresponding author*: ravindrabhardwaj2@gmail.com

Rajesh Kumar Chakrawarti, Ranjana Sikarwar, Sanjaya Kumar Sarangi, Samson Arun Raj Albert Raj, Shweta Gupta, Krishnan Sakthidasan Sankaran and Romil Rawat (eds.) Natural Language Processing for Software Engineering, (1–14) © 2025 Scrivener Publishing LLC

data and networks. Conventional intrusion detection technologies are ineffective in detecting security breaches in the Internet of Things (IoT) because of the distinct standards and protocol stacks used in its network. Regularly analyzing the vast amount of data created by IoT is a tough task due to its endless nature. An intrusion detection system (IDS) is employed to safeguard a system or network against unauthorized access by actively monitoring and identifying any potentially malicious or suspicious activities. Machine learning technologies provide robust and efficient approaches for mitigating these distinct hazards. The establishment of a robust machine learning system is the key to acquiring networks that are free from any form of threats.

***Keywords*:** Machine learning, Internet of Things, security, privacy, attacks, vulnerability, intrusions

1.1 Introduction

The use of connected devices made ordinary chores easier and more efficient. They also provide a lot of information that is of great use. Connected automobiles, for example, may be able to take use of services that provide driver assistance. Medical devices give detailed patient records. The unfortunate reality is that a digital assault is possible on any device that is capable of establishing a connection to the internet. In worst case, many of these devices are missing even the most basic safety safeguards. According to the authors of the report, almost all of the data flow associated with the internet of things (98%) is not secured. This information may be obtained by anybody with little effort. To repeat, devices that are connected to the Internet of Things provide fraudsters with an easy target. Not only might their information be stolen, but perhaps other sensitive data as well. Using one of these devices is a frequent strategy used by hackers to gain access to a company's internal network. The sheer number of these devices and the settings they control may be enough to pique the interest of a cyber-attacker [1] as given in Figure 1.1: Increasing Number of DDOS Attacks [Source: Cisco Annual Internet Report 2018-2023] and in Figure 1.2: Threats to Internet of Things.

In a smart environment, any number of items, including databases of user credentials, electronic sensors, CCTV installations, access controls, personal electronic devices, recorded biometrics, and so on, might be the target of an attack. It is essential to protect the confidentiality, integrity, availability, authentication, and authorization features of the IoT architecture from a security point of view [2]. DDoS attacks are becoming more common, and Cisco's Annual Internet Report (2018-2023) White Paper forecasts that the total number of DDoS attacks would more than double

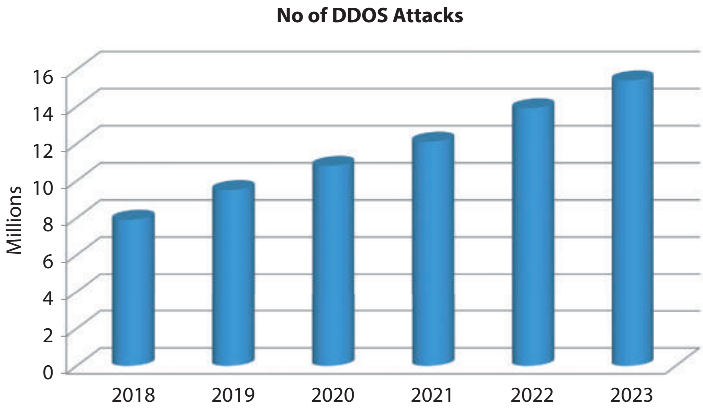


Figure 1.1 Increasing number of DDOS attacks [Source: Cisco Annual Internet Report 2018-2023].

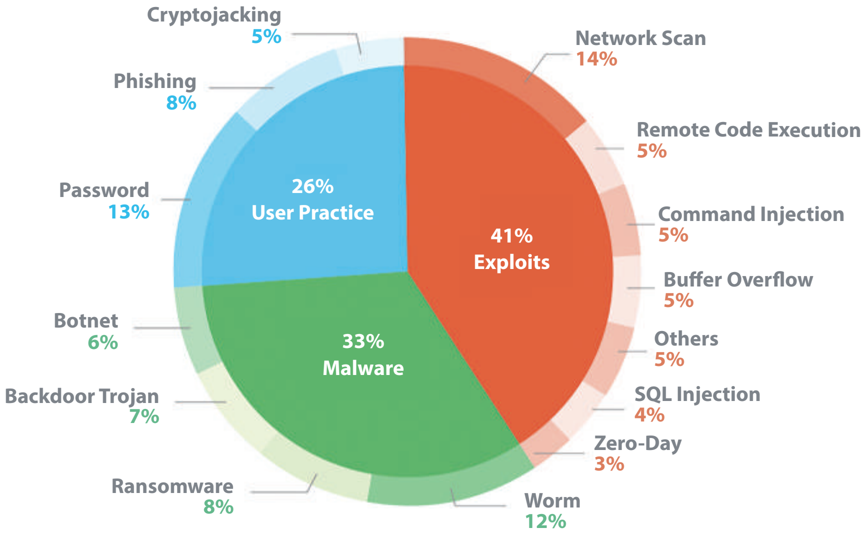


Figure 1.2 Threats to Internet of Things.

from the 7.9 million that were seen in 2018 to anywhere over 15 million by 2023 as shown in Figure 1.1.

According to the survey, 57% of IoT devices that are connected via this insecure traffic are susceptible to medium- to high-severity attacks, making them an easy target for cybercriminals [3]. In addition, the survey found that 41% of attacks target IoT vulnerabilities by scanning them against publicly available databases of known security flaws. The analysis is shown in Figure 1.2.

According to the Internet of Things Threat Report published by Palo Alto Networks in March 2020, 98% of all traffic from IoT devices is unencrypted, giving attackers a chance to eavesdrop. This network contains sensitive and private information that is easily accessible to attackers, who may then sell the information on the dark web for a profit.

1.2 Need of Vulnerability Identification

Vulnerabilities in IoT network are increasing every year. As shown in Figure 1.3, IoT environment is experiencing, a large number of new vulnerabilities every year. All the Internet of Things applications—smart city, smart farming, smart healthcare, smart transportation, and smart traffic—are experiencing new vulnerabilities and increasing number of attacks every year. Also, vulnerabilities and attacks are increasing every year. Number of vulnerabilities has increased threefold in the last decade and twofold in last five years as represented in Figure 1.3: Number of New Vulnerabilities Identified in IOT [Source- IBM X-Force Threat Intelligence Index 2022].

The process of determining how vulnerable a system is to attack is referred to as a vulnerability scan. This kind of scan is carried out to identify potential entry points into a computer or network so that appropriate preventative measures may be taken. Automated scanning methods check applications to see if they have any security problems to establish whether

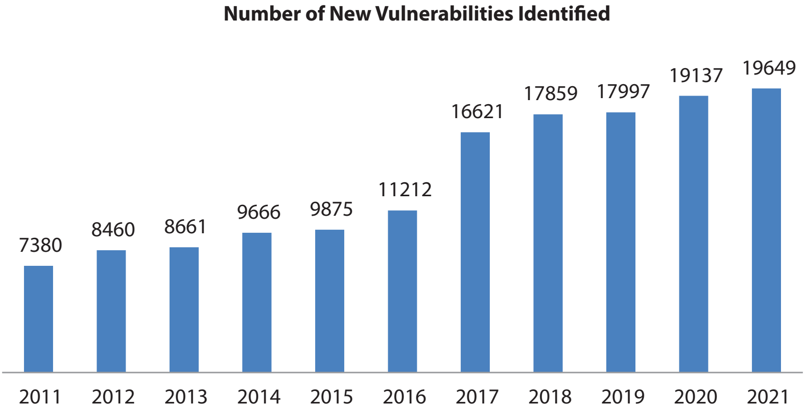


Figure 1.3 Number of new vulnerabilities identified in IoT [Source- IBM X-Force Threat Intelligence Index 2022].

or not there are vulnerabilities in an organization's internal network. Users are spared the time and effort required to carry out hundreds or even thousands of manual tests for each kind of vulnerability since vulnerability scanners automate the process of searching for security issues in a system.

To maintain the integrity of the system's protections, it is essential to assign vulnerabilities a severity ranking before putting into action any remedial procedures. Common Vulnerability Scoring System (CVSS) is a tool that administrators may use to prioritize security problems according to the severity level associated with each fault. The CVSS score of vulnerability is a standard metric that is not developed for unique network architecture. Despite the fact that the frequency and impact of vulnerabilities affect the security risk level of a specific network, the CVSS score of vulnerability is a standard metric. In addition to the severity score, a number of other factors also affect the level of security risk that is posed by the organization's underlying infrastructure. These factors include the age and frequency of vulnerabilities already present in the system, as well as the impact that exploiting vulnerability has on the system. For this reason, it is advised that, when doing risk level calculations, these components, together with the CVSS severity score, be used. This will allow for effective network security risk management.

1.3 Vulnerabilities in IoT Web Applications

The authors of [4] provide a code inspection-based strategy. To identify a number of mistakes hidden inside the process, this method makes use of code inspection. It is said that the offered approach may be used to locate each and every vulnerability in the NVD. Using this classifier might assist in more accurately identifying potential security flaws.

In addition, a web crawler was developed by Guojun and his colleagues [5]. This web spider collects papers that are connected to one another. The TF-IDF is essential to the methodology. Medeiros *et al.* [6] were the ones who first proposed the approach for evaluating the quality of the code. The concepts that underlie data mining are built on this methodology, which acts as the basis for those concepts. New techniques for identifying web server vulnerabilities were developed by [7].

Authors [8] have developed an innovative method for locating vulnerabilities in web applications. In addition to this, static analysis and data mining directly from the source code are used. Researchers [9] came to the conclusion that XML injection is a critical issue that exists in all web applications.

The vast majority of recently published web apps continue to be plagued by XML injection difficulties.

According to research by [10], a large percentage of such norms rely on online application security. Security measures designed to prevent code injection attacks on web applications were the primary focus of these studies. But even if the notion of acceptance is clearly defined and extensively concealed in almost all international standard regulations, the number of assaults is rising because of flaws in the infusion of code. This is the opinion of the developers. To reduce safety gauges, it is crucial to inform engineers and clients about the relevance of these metrics and to urge them to fulfil the standards with meticulous care. The time we waste waiting for this type of instruction and support is just not acceptable.

Authors [11] spoke about the significant factors that are engaged in the life cycle of product innovation. In addition, a number of software engineers have introduced security mechanization tools and processes that can be used at any stage of the software development life cycle (SDLC) to enhance the stability and quality of even the most fundamental digital systems. In addition to this, they requested that all organizations working to improve networks place a higher priority on planning, education, risk assessment, threat modelling, audits of architecture configuration, secure coding, and assessments of data that has been sent and received after it has been processed.

Wang and Reiter [12] developed a method for mitigating denial of service attacks by making use of a website's diagrammatic structure to counter flooding assaults. When visiting the destination website, a valid customer has the opportunity to quickly get a reward URL by clicking on a referral link provided by a reputable source. The proposed paradigm has no requirements in terms of infrastructure, and it does not call for any changes to be made to the code that users use when they access websites. The WRAPS framework, in addition to the intentions that its creator had for it, was provided. Nearly all of the smart assaults on websites recycled old strategies and methods from earlier attacks. There is a wide number of guises under which one may launch an assault against a strategy or an approach. They may also be seen in circumstances that are not related to the web. Attacks on a website's business logic may be harmful to the website itself, but attackers can also utilize websites as a go-between to accomplish their goals.

The SQLProb [13] will remove the user input and check to see whether it complies with the syntactic requirements of the query. This is accomplished by applying the formula that was inherited and then improving it. The SQLProb is a comprehensive discovery approach that does not need any modifications to be made to either the application or the database. This allows it to avoid the complexity of polluting, learning, and instrumenting code. In addition, neither education nor metadata are required in order to go on with the material's approval procedure.

Authors presented a complete stream-based WS-security handling architecture in their paper [14]. This design improves the level of preparedness in the administration processing and raises the level of resistance to different kinds of DoS assaults. When leaking is used as a strategy, their engine is able to handle standard WS-Security application scenarios.

The author [15] has examined the vast majority of the conventional criteria that are used to judge Web service quality. The majority of the measures, including performance, consistency, adaptability, limit, strength, exception handling, correctness, uprightness, openness, accessibility, interoperability, and security, all fall below the average level.

Hoquea *et al.* [16] took into consideration the activities that may be taken as well as the probable results or degrees of harm. Following that, the designer divides the assaults into a number of distinct categories. They consistently offered a scientific classification of attack equipment to assist in the organization of security specialists. This was done to help in the prevention of potential threats. They delivered a detailed and well-organized examination of existing tools and frameworks that may aid attackers as well as system defenders. Their focus was on tools and frameworks that are available now. The writers have included a description of both the benefits and drawbacks of the tools and frameworks in the event that you are interested in learning more about them.

Binbin Qu *et al.* [17] provided an explanation of the method that lies behind a model design. The construction of a pollutant dependency diagram for the program requires many steps, one of which is a static examination of the program's source code. They employ a limited state automaton to adhere to the attack model while communicating the pollutant string estimate and verifying the robustness of the program's protections for user input. All of this takes place while maintaining the integrity of the attack model. They utilized the framework model for computerized recognition based on the examination of the spoils and placed it into operation.

1.4 Intrusion Detection System

An incursion refers to any malevolent or dubious activity that jeopardizes the security of a computer or network. Intruders may originate from either internal or external sources. Internal intruders conceal themselves within the targeted network and acquire elevated privileges to deliberately harm the network infrastructure. External intruders surreptitiously extract data from the target network while remaining concealed outside of it. Internal attacks are initiated by nodes that are either malevolent or compromised, whereas external assaults are initiated by entities that are external to the system. An intrusion detection system (IDS) refers to any hardware or software that can identify and alert to potentially malicious activity on a network or computer system. Moreover, it may also be employed to detect any dubious activities or breaches within the system. Typically, when a network or system behaves abnormally, it suggests the occurrence of anything violent, harmful, or illegal. Although the majority of intrusion detection systems (IDS) mostly depend on identifying and reporting anomalies, there are a handful that excel in detecting intrusions that are overlooked by conventional firewalls. In terms of safeguarding the system from harm, intrusion detection systems (IDS) function similarly to firewalls by preventing unauthorized individuals from gaining access.

There are a total of three categories of intrusion detection systems based on the source of data, four groups based on the technique of analysis, and an additional three groups in total.

The Host-Based Intrusion Detection System (HIDS) software is placed on a computer to monitor, evaluate, and gather data on the traffic and suspicious activities of that specific system. In addition, it analyses not just the traffic activity, but also the system calls, file system changes, inter-process communication, and program running on the computer (ZarpelÍo *et al.*, 2017). HIDS utilizes data collected from the operating system and application software to detect suspicious activities. When a host-based intrusion detection system (HIDS) is deployed, it is capable of detecting intrusions solely on the host where it is installed. Installation of HIDS eliminates the need for extra software to identify threats on the system. Intruder detection systems are designed to detect and identify instances of unauthorized access or attacks from within a protected area. The installation cost is substantial due to the requirement of individual Host-based Intrusion Detection Systems (HIDS) for each device as given in Figure 1.4: Host-based IDS.

The Network-Based Intrusion Detection System (NIDS) safeguards network nodes by capturing and scrutinizing all network packets for malicious activities. Figure 1.5 displays the structure of the NIDS. The sensor is strategically positioned in a vulnerable region inside the

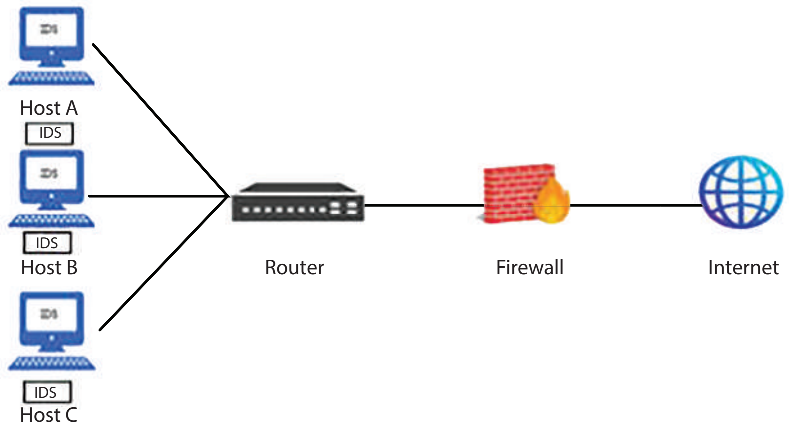


Figure 1.4 Host-based IDS.

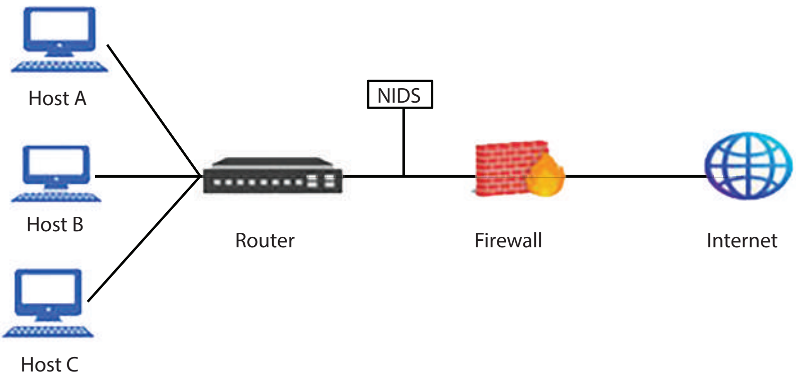


Figure 1.5 Network-based intrusion detection system.

network, bridging the server and the network. The NIDS monitors both incoming and outgoing communications. If the system identifies any network risks, it will need to respond rigorously in order to safeguard itself. One possible course of action is to prohibit network access from the specified IP address, while another alternative is to inform the responsible party through warning notifications. Determining if the NIDS has noticed their potential intrusions might provide a challenge for a thief. Monitoring extensive networks is under the purview of only a limited number of intrusion detection systems. To mitigate potential security risks, it is imperative to implement scanners, sniffers, and network intrusion detection tools. These measures are necessary to safeguard against various malicious activities such as IP spoofing, DOS assaults, DNS name corruption, man-in-the-middle attacks, and arp cache poisoning. These vulnerabilities arise due to the inherent weaknesses in TCP/IP protocols represented in Figure 1.5 Network-Based Intrusion Detection System.

Hybrid Intrusion Detection Systems (HIDS) integrate the functionalities of several intrusion detection systems to identify and expose intrusions. A hybrid intrusion detection system integrates data from both the network and the host agent or system to create a full overview of the network system. The hybrid technique is the most effective strategy for intrusion detection. Prelude is an example of a hybrid intrusion detection system.

1.5 Machine Learning in Intrusion Detection System

Soft computing makes it possible to build intelligent machines that are able to solve challenging issues that arise in the real world but are beyond the purview of standard mathematical modelling. These kinds of problems cannot be adequately modelled using traditional methods. It has a high tolerance for approximate information, ambiguity, imprecision, and merely a partial view of the environment [18], which enables it to emulate the way individuals form their opinions and make decisions. In this section, we will have a brief discussion on the many different techniques to soft computing that may be used in the process of detecting intrusions.

The genetic algorithm (GA) is a search engine that has been in use since it was conceived in Holland. This search engine is both strong and adaptable. There it first emerged in its current shape for the first time. Because of advances in technology, it is now possible to recreate the natural process of evolution that takes place in uncontrolled environments. The GA may be seen in this way as an example of a global search process that depends on randomness. The concept of "survival of the fittest" is applied by the algorithm to the challenge of developing ever more accurate approximations of a solution to the issue.

The most experienced people in the sector are recruited to teach the next generation, which ultimately results in the development of novel solutions to the issue. If this approach is used, the newly recruited staff members could be better able to address the current challenge [19]. The fitness function enables us to get insight into how well people fared on the aspects of the exam that were the most challenging [20].

PSO was first developed in 1995 by [21], who drew their inspiration from the way fish and birds congregate in groups known respectively as flocks and schools. In an effort to discover a solution, a "population" of particles is moved over the damaged region at specified speeds and rotated clockwise and anticlockwise. By employing the stochastic calibration approach and taking into consideration the best preceding and best adjacent locations of the particles, the velocities of the particles may be changed appropriately. A random number generator is what's needed to get this done.

A kind of logic known as fuzzy logic is one that employs the practice of approximation. The paradigms for optimization and classification used in machine learning are both underpinned by evolutionary computing, which is based on genetic and natural selection-based evolutionary processes. The origin of these evolutionary processes may be traced back to evolution. The majority of the time, genetic algorithms are used [22] in applications that are based on the actual world of business.

In contrast to the conventional naive Bayesian classifier, the HNB may take on a variety of forms depending on the circumstances. Finding the attribute's hidden parent needs the inclusion of a further layer in the HNB model, which necessitates the addition of this layer. The structures of the HNB components may be inferred with the help of Naive Bayes. Each characteristic has a hidden past that was fostered to bring together the many energy that it symbolizes. For the purpose of providing an overview of the covert parents, we may make use of the mean of weighted one-dependency estimators [23, 24].

The support vector machine, sometimes known as an SVM [25, 26] for short, is a technique to classification that is grounded on statistical learning theory (SLT) [27–29]. Another kind of system that is comparable is known as a hyper-plane classifier. In support vector machines (SVM), a good hyper-plane is one that successfully separates the classes while keeping the amount of interclass overlap to a bare minimum.

Deep neural networks, more often referred to as DBNs, are generative graph models that are used in machine learning. These networks are built on latent variables, which are also referred to as hidden units. These networks simply link the levels themselves, and not the units that are included inside those levels.

We may look at the model that was built by researchers and published in [24] as an illustration of one method that can be used to determine attributes for an intrusion detection system.

1.6 Conclusion

The issue of safety is of utmost importance in the context of IoT and other types of pervasive connectivity. There is a growing probability that attacks would focus on companies and organizations that utilize IoT. Traditional cybersecurity systems face multiple obstacles when attempting to detect zero-day threats. The invader exploits the privileges offered by the IoT architecture to acquire valuable data. There are few security risks that are widely recognized, and even fewer that involve slow and unnoticed attacks. An effective strategy to tackle these unexpected challenges is to construct intrusion detection systems using machine learning techniques. Cyberattacks on the Internet of Things architecture may result in data loss or information loss, as well as IoT device sluggishness. To guarantee the security of data and networks, intrusion detection systems have been in use for the last 20 years. Because the Internet of Things (IoT) uses unique standards and protocol stacks, traditional intrusion detection methods are not successful in identifying security breaches in its network. Because the amount of data generated by IoT is infinite, it is difficult to regularly analyze it. A system or network is protected from unauthorized access by an intrusion detection system (IDS), which actively monitors and detects any potentially harmful or suspicious activity. Machine learning technologies offer reliable and effective methods for reducing these specific risks.

References

1. Raghuvanshi, A., Singh, U.K. *et al.*, Intrusion Detection Using Machine Learning for Risk Mitigation in IoT-Enabled Smart Irrigation in Smart Farming. *J. Food Qual.*, 2022, 1, 1–8, 2022.
2. Abhishek, R., Singh, U.K., Phasinam, K., Kassanuk, T., Internet of Things-Security Vulnerabilities and Countermeasures. *ECS Trans.*, 107, 1, 15043–15053, 2022.
3. Raghuvanshi, A., Singh, U.K., Joshi, C., A Review of Various Security and Privacy Innovations for IoT Applications in Healthcare. *Adv. Healthcare Syst.*, 1, 43–58, 2022, doi: 10.1002/9781119769293.ch4.
4. Zhang, Q. and Wang, X., SQL injections through back-end of RFID system, in: *2009 International Symposium on Computer Network and Multimedia Technology. CNMT 2009*, pp. 1–4, IEEE, 2009.
5. Li, Z. *et al.*, VulPecker: an automated vulnerability detection system based on code similarity analysis. *ACM, Proc. of the 32 Annual Conference on Computer Security Applications*, p. 201213, 2016.

6. Guojun, Z. *et al.*, Design and application of intelligent dynamic crawler for web data mining, in: *Automation (YAC), 2017 32nd Youth Academic Annual Conference of Chinese Association*, pp. 1098–1105, IEEE, 2017.
7. Medeiros, I., Neves, N., Correia, M., Detecting and removing web application vulnerabilities with static analysis and data mining. *IEEE Trans. Reliab.*, 1, 54–69, IEEE, 2016.
8. Masood, A. and Java, J., Static Analysis for Web Service Security – Tools & Techniques for a Secure Development Life Cycle. *International Symposium on Technologies for Homeland Security*, pp. 1–6, 2015.
9. Medeiros, I. and Neves, N., Detecting and Removing Web Application Vulnerabilities with Static Analysis and Data Mining. *IEEE Trans. Reliab.*, 1, 1–16, 2015.
10. Salas, M.I., de Geus, P.L., Martins, E., Security Testing Methodology for Evaluation of Web Services Robustness - Case: XMLInjection. *IEEE World Congress on Services*, pp. 303–310, 2015.
11. Madan, S., Security Standards Perspective to Fortify Web Database Applications from Code Injection Attacks. *International Conference on Intelligent Systems, Modelling and Simulation*, pp. 226–233, 2010.
12. Teodoro, N. and Serrao, C., Web application security: Improving critical web - based applications quality through in - depth security analysis, in: *International Conference on Information Society (i- Society)*, pp. 457–462, 2011.
13. Wang, X. and Reiter, M.K., Using Web-Referral Architectures to Mitigate Denial-of-Service Threats. *J. IEEE Trans. Dependable Secure Comput.*, 7, 2, 203–216, 2010.
14. Liu, A., Yuan, Y., Wijesekera, D., Stavrou, A., SQLProb: a proxy-based architecture towards preventing SQL injection attacks, in: *Proceedings ACM Symposium on Applied Computing (SAC'09)*, pp. 2054–2061, 2009.
15. Gruschka, N., Jensen, M., Lo Iacono, L., Luttenberger, Server-side Streaming Processing of WS-Security. *IEEE Trans. Serv. Comput.*, 4, 4, 272–285, 2011.
16. Ladan, M.I., Web Services Metrics: A Survey and A Classification. *J. Commun. Comput.*, 9, 7, 824–829, 2012.
17. Hoque, N., Bhuyan, M.H., Baishya, R.C., Bhattacharyya, D.K., Kalita, Network Attacks: Taxonomy, tools and systems. *J. Comput. Netw. Appl.*, 1, 13–26, 4 October 2013, doi: doi.org/10.1016/j.jnca.2013.08.001.
18. Kulshestha, G., Agarwal, A., Mittal, A., Sahoo, A., Hybrid Cuckoo Search Algorithm for Simultaneous Feature and Classifier Selection. *IEEE International Conference on Cognitive Computing and Information Processing (CCIP)*, pp. 1–6, 2015.
19. Visumathi, J. and Shunmuganathan, K.L., A computational intelligence for evaluation of intrusion detection system. *Indian J. Sci. Technol.*, 4, 1, 28–34, Jan 2011.
20. Wang, B., Yao, X., Jiang, Y., Sun, C., Shabaz, M., Design of a Real-Time Monitoring System for Smoke and Dust in Thermal Power Plants Based

on Improved Genetic Algorithm. *J. Healthc. Eng*, 2021, D. Singh (Ed.), pp. 1–10, Hindawi Limited, UAE, 2021, https://doi.org/10.1155/2021/7212567.

21. Mohanasundaram, S., Ramirez-Asis, E., Quispe-Talla, A., Bhatt, M.W., Shabaz, M., Experimental replacement of hops by mango in beer: production and comparison of total phenolics, flavonoids, minerals, carbohydrates, proteins and toxic substances, *Int. J. Syst. Assur. Eng. Manage.*, Springer Science and Business Media LLC, UAE, 2021, https://doi.org/10.1007/s13198-021-01308-3.
22. Almahirah, M.S., S, V.N., Jahan, M., Sharma, S., Kumar, S., Role of Market Microstructure in Maintaining Economic Development. *Empirical Econ. Lett.*, 20, 2, 01–14, 2021.
23. Chaudhary, A., Tiwari, V.N., Kumar, A., Analysis of Fuzzy Logic Based Intrusion Detection Systems in Mobile Ad Hoc Networks. *Int. J. Inf. Technol.*, 6, 1, 183–198, June 2014.
24. Rathore, N. and Rajavat, A., Smart Farming Based on IOT-Edge Computing: Applying Machine Learning Models For Disease And Irrigation Water Requirement Prediction In Potato Crop Using Containerized Microservices, in: *Precision Agriculture for Sustainability*, pp. 399–424, Apple Academic Press, UAE, 2024.
25. Patsariya, M. and Rajavat, A., A Progressive Design of MANET Security Protocol for Reliable and Secure Communication. *Int. J. Intell. Syst. Appl. Eng.*, *12*, 9s, 190–204, 2024.
26. Rathi, M. and Rajavat, A., Investigations and Design of Privacy-Preserving Data Mining Technique for Secure Data Publishing. *Int. J. Intell. Syst. Appl. Eng.*, *11*, 9s, 351–367, 2023.
27. Dubey, P. and Rajavat, A., Effective K-means clustering algorithm for efficient data mining, in: *2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN)*, pp. 1–6, IEEE, 2023, May.
28. Nahar, S., Pithawa, D., Bhardwaj, V., Rawat, R., Rawat, A., Pachlasiya, K., Quantum Technology for Military Applications. *Quantum Comput. Cybersecur.*, 1, 313–334, 2023.
29. Pithawa, D., Nahar, S., Bhardwaj, V., Rawat, R., Dronawat, R., Rawat, A., Quantum Computing Technological Design Along with Its Dark Side. *Quantum Comput. Cybersecur.*, 1, 295–312, 2023.

2

Frequent Pattern Mining Using Artificial Intelligence and Machine Learning

R. Deepika[1]*, Sreenivasulu Gogula[2], K. Kanagalakshmi[3], Anshu Mehta[4], S. J. Vivekanandan[5] and D. Vetrithangam[6]

[1]Department of AI&DS, B V Raju Institute of Technology, Narsapur, Telangana, India
[2]Department of CSE (Data Science), Vardhaman College of Engineering, Shamshabad, Hyderabad, India
[3]Department of Computer Applications, SRM Institute of Science and Technology (Deemed to be University), Trichy, India
[4]Department of Computer Science and Engineering, Chandigarh University, Mohali, Punjab, India
[5]Department of Computer Science and Engineering, Dhanalakshmi College of Engineering, Dr V P R Nagar, Manimangalam, Tambaram, Chennai, India
[6]Department of Computer Science & Engineering, University Institute of Engineering, Chandigarh University, Mohali, Punjab, India

Abstract

Frequent pattern mining is a very active topic in the field of data mining. Numerous researchers have considered it since its beginning. The dimensions of all areas expand exponentially with the advancement and accumulation of data. The ability to effectively and easily assess and extract time-sensitive information from large datasets is essential for making informed decisions and uncovering new knowledge. Data mining is the use of sophisticated analytics on large databases to discover previously unidentified links, patterns, and trends. Efficient and adaptable handling of large-scale data is crucial for retrieving information and making informed decisions. Data mining is the systematic analysis of vast quantities of data to uncover previously undiscovered correlations, patterns, and trends. Since the inception of the World Wide Web, there has been a rapid and significant increase in the quantity of data that is stored and can be accessed electronically. Data mining, which refers to the process of extracting new insights from data, has

**Corresponding author*: deepika.r@bvrit.ac.in

Rajesh Kumar Chakrawarti, Ranjana Sikarwar, Sanjaya Kumar Sarangi, Samson Arun Raj Albert Raj, Shweta Gupta, Krishnan Sakthidasan Sankaran and Romil Rawat (eds.) Natural Language Processing for Software Engineering, (15–28) © 2025 Scrivener Publishing LLC

become a crucial tool for both business and academic sectors. With the introduction of the Internet, there has been a rapid increase in the quantity of data stored and available online. Consequently, the methods for extracting valuable information from this extensive collection of data have become crucially significant in several domains, such as business and academics. Frequent Item Set Mining is a very popular technique for getting significant insights from datasets.

***Keywords*:** Frequent pattern mining, decision tree, KNN, accuracy, machine learning

2.1 Introduction

Data mining is the systematic exploration of extensive databases to discover noteworthy and previously unidentified patterns [1]. The step described by Fayyad *et al*. for Knowledge Discovery in Databases (KDD) is included in this approach. Data cleaning, integration, selection, transformation, mining, pattern assessment, and knowledge representation are all integral components of the continuing process known as Knowledge Discovery in Databases (KDD). Data mining may be used to several types of data. The approaches and procedures may vary when used to different sorts of data. The patterns extracted from data might vary in terms of their nature and the specific sort of data mining task. Data mining jobs may be broadly classified into two categories: descriptive and predictive. Predictive data mining utilizes existing data to generate predictions, whereas descriptive data mining aims to explain the overall characteristics of the provided data.

Bayes' hypothesis and relapse inquiry were employed in the 1700s to distinguish designs from noise (1800s). Increases in PC innovation have led to a broader variety and higher capacity for information. Hands-on information examination has expanded as the quantity and complexity of informative indexes have grown. There have been a variety of software engineering breakthroughs that have led to this progress, such as the discoveries of neural systems, bunching, hereditary computations (1950s), decision trees (1960s), and support vector machines (1980s) [2].

In a decision tree, each node represents an evaluation of some attribute's value, and each branch represents the evaluation's outcome. The tree's leaves represent classes or distributions of classes. It's a cinch to convert from decision trees to characterization rules. Decision trees can cope with a lot of information. When it comes to storing data, they use a tree structure that is intuitive and simple to learn. Using a decision tree is a straightforward process that requires just a few easy steps to understand and put together. Decision tree enlistment computations have been used in a wide range of fields, including medical, manufacturing, budgeting, cosmology, and subatomic research [3].

Both AI and data mining heavily rely on tree-based learning approaches. It's no secret that these strategies have been in use for a long time. There is nothing over the top about them, and that's exactly what makes them so endearing. When making decision trees, a top-down strategy is often used to identify a univariate split that boosts some local basis (for example, gain percentage) until the leaf segments of the tree are sufficiently pure. Pessimistic Error Pruning uses heuristics that may be measured, while Reduced Error Pruning uses a single set of pruning to determine this utility.

It is a very costly strategy to employ the Naive Bayes classifiers as leaf hubs in all of the first-level child hubs (evaluated by cross-approval), yet this is the only way NB Tree can deliver them in a decision tree. At each node, students analyze additional characteristics as straight, quadratic, or calculated attribute elements, and these elements are then sent down the tree in the same manner that they were processed before. However, despite the fact that root-to-leaf probability dispersions are referred to as disseminations, leaf hubs remain the primary classifiers [4].

This study introduces a recursive Bayesian classifier. One hundred percent accuracy in decision tree enlistment has previously been achieved by a variety of methods, and many of them have been successful. As a result, these new approaches were time-consuming and difficult to learn, and this was the major issue. Recursively dividing the data into places where there is a suspicion of constraining freedom is the most significant aspect. Planning from perceptions of the item to choices based on those perceptions is how judgements about the objective value of anything are made [5].

Determining whether a system is well on its way to attaining its objective is the most prominent usage of decision trees in tasks research. Restrictive probabilities may be calculated using decision trees. A decision tree (also known as a tree outline) is a decision aid that employs a tree-like diagram or model to describe alternatives and their probable outcomes, such as chance event effects, asset expenditures and utility. The decision tree induction approach has been used effectively in master frameworks to gather information. It is possible to use decision trees to enrol people from a variety of data sources [6].

2.2 Data Mining Functions

Information mining is an assortment of procedures for proficient computerized disclosure of beforehand obscure, substantial, novel, helpful, and reasonable examples in enormous databases. The examples must be significant with the goal that they might be utilized in an endeavor's dynamic procedure [7]. Information mining procedures can be gathered as follows as given in Figure 2.1: Data Mining Methods:

- Classification-It is necessary to classify the supplied information event into one of the objective classes that have already been identified or defined. One of the models may be whether a customer is a trustworthy client or a defaulter in Visa's interchange information base, based on his distinct segment and previous purchase characteristics [8].
- Estimation-Like order, the motivation behind an estimation model is to decide an incentive for an obscure yield trait. In any case, in contrast to grouping, the yield quality for an estimation issue is numeric as opposed to clear cut.
- Prediction-It isn't anything but difficult to separate forecast from grouping or estimation.

The primary distinction lies in the fact that the predictive model extrapolates results into the foreseeable future rather than providing directives

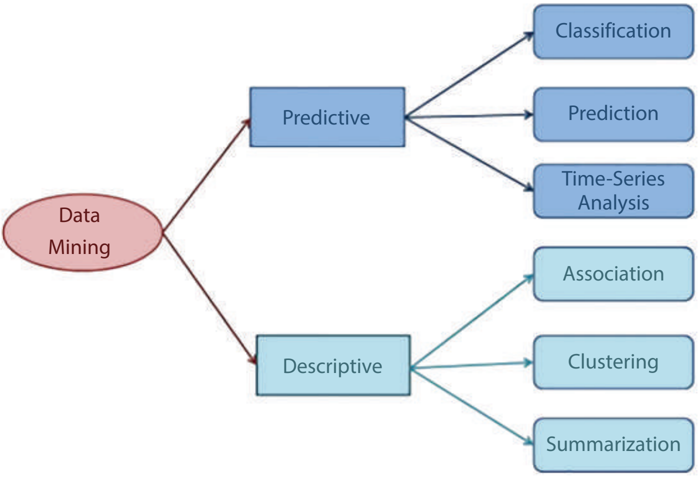


Figure 2.1 Data mining methods.

for actions in the here and now. The discontinuous or quantitative nature of the output characteristic can be chosen. One illustration of what a model might entail is making a forecast regarding the value of the Dow Jones Industrial Average at the end of the following week, and explains the history of a decision tree as well as its possible applications in more detail.

- Association rule mining-Here interesting hidden rules called affiliation rules in a huge value-based information base is mined out. For example, the standard {milk, margarine >biscuit} gives the data that at whatever point milk and spread are bought together scone is additionally bought, with the end goal that these things can be set together for deals to build the general deals of every one of the things [9].
- Clustering-Clustering is an uncommon kind of grouping where the objective classes are obscure. For example, given 100 clients they must be characterized depending on certain comparability measures and it isn't biased [10].

The fundamental application zones of information mining are in Business investigation, Bioinformatics, Web information examination, text investigation, sociology issues, biometric information examination and numerous different spaces where there is extension for shrouded data recovery. A portion of the difficulties before the information mining analysts are the treatment of unpredictable and voluminous information, conveyed information mining, overseeing high dimensional information, and model enhancement issues [11].

In the coming areas, the different stages happening in an ordinary information mining issue are clarified. The different information mining models that are usually applied to different issue spaces are additionally examined in detail in the coming areas.

2.3 Related Work

Information readiness or pre-preparing is a significant advance where the information is made reasonable for handling. This includes cleaning information, information changes, choosing subsets of records, and so on. During the process of setting up one's understanding, there are two stages: determination and change. Methods of deliberation for the collection of information can be applied to the process of data extraction. To accomplish this work, one will first need to perform a search within the database to retrieve the desired characteristics. The process of converting unstructured, unprocessed data into a shape that can be processed effectively by an information management or storage system is referred to as "data transformation." Another name for this process is "data articulation."; e.g., symbolic information types are changed over into numerical structure or straight-out structure [12].

Models such as decision trees, neural systems, and so on are built using the above-mentioned information. To uncover hidden patterns in data, information mining is a method of applying various tactics to data. Finding several models (or capabilities) that describe and identify information classes or concepts is the task of orderly management. An extensive amount of preliminary data was analyzed to derive the inferred model (i.e., information protests whose class mark is known). Different structures may be used to communicate with the inferred model, such as IF-THEN rules, decision trees, scientific formulas, or neural networks.

Grouping techniques like the decision tree are well-known. Tree-like structure with each hub represents a choice on one of the following values: the tree's branches represent the decisions made, while its leaves represent the many classes. A decision tree is both predictive and expressive. Using a decision tree, one can see how the information that were gathered connects.

A decision tree is a useful instrument for planning in the fields of data mining and artificial intelligence. It allows one to shift from more subjective evaluations of an object to more objective evaluations of the object's value. These tree models are known by a few different names, including the layout tree for the purpose of analyzing a particular outcome and the regression tree for the purpose of analyzing repetition, to name just two examples. (ceaseless result). The branches and greenery of the trees are meant to represent the relationships that exist between the various components that come together to form the various organizations. Data is used in an approach to artificial intelligence known as decision tree learning, which results in the generation of a decision tree.

The purpose of the computation is to become familiar with the judgment capacity that is stored in the information and then use it to organize new sources of data given a large number of models (preparing information) represented by some organization of characteristics. The objective of the computation is to become familiar with the judgment capacity that is stored in the information. (for example, sex, position, foundation). One can determine the discriminative strength of each characteristic that can most effectively separate the dataset by applying either the idea of data addition or the Gini list. Both of these methods are described in the following sentence. While common programs such as CART and IBM's intelligent digger make use of the Gini list concept, well-known decision tree algorithms such as ID3, C4.5, C5, and so on depend on data development to determine the next best property [13].

The algorithm determines the characteristics of the first dataset, which are the p-value for the number of positive models and the n-value for the number of negative models. After that, the sample is divided along the chosen characteristic (by making use of v-decisions), and the information gain is computed. This process is replicated for each individual piece of real estate until the location that provides the most significant increase in data gain is selected to serve as the primary nerve center. There are surprisingly few distinctions between ID3, C4.5, and C5 and other similar formats, despite the fact that they all use data gain principles. The techniques that are used to generate and present decision trees are broken down and analyzed in this section.

Continuous and discontinuous feature management is possible with C4.5. To deal with irreversible characteristics, C4.5 first establishes a limit; and then, it separates the list into those whose attribute respect is higher than the limit, and those whose attribute respect is not exactly or exactly equal to the limit. When assembling data with missing quality attributes, it is beneficial to have C4.5's missing attribute support because it enables designate qualities to be distinguished as "?" for missing. The lacking property approximations are not used in either the addition or the entropy calculations. The notion of clipping is utilized by C4.5. After the tree has been constructed, the algorithm will iteratively traverse it to prune away insignificant branches and replace them with leaf nodes where appropriate. C5.0 is a substantial enhancement over C4.5 in terms of both the speed at which it operates and the effectiveness with which it uses memory. Additionally, it makes use of forward-thinking ideas such as bolstering, which are broken down into their component parts in later portions of the theory.

Calculations based on the data gain theory will be more charitable to characteristics that have a greater number of qualities as the number of target groups increases, whereas calculations based on the Gini list will be less generous in this regard. Therefore, contemporary researchers are making an effort to improve decision tree exhibits by utilizing techniques such as pre-pruning (the removal of unnecessary tree nodes during tree assembly) and post-pruning (the removal of tree nodes after tree construction), and others as well.

Computations involving decision trees can be performed with CS4 in a variety of ways, including Bagging, Boosting, and Random Backwoods. Within the framework of the cross-approval set, a hub is only eliminated if the subsequent reduced tree produces results that are comparable to the initial one. Because the presentation is based approximatively on an approval set, this technique of tree pruning has the drawback that the actual tree can only make use of a smaller amount of information. However, over time, C4.5 will generate a certain amount of error based on the act of preparing the information itself. This error will be calculated using the upper limit of a confidence stretch, which is generally 25%, and will be applied to the re-replacement error. The estimated within one standard deviation of the anticipated error of the wheel. The additional cutting function known as sub tree rising in C4.5 provides a number of different options, including reduced mistake clipping as one of them. In the event that a sub tree is elevated, a lower hub might take the position of a higher hub, and vice versa; this would cause a reorganization of the tests.

A point-by-point outline on how C4.5 conducts its post-pruning is given in [14]. Different calculations for decision tree acceptance incorporate ID3 (antecedent of C4.5), C5.0 (replacement of C4.5), CART (grouping and relapse trees) [8], LMDT (Linear Machine Decision Trees) [15], OC1 (slanted classifier, etc. as given in Figure 2.2: A Sample decision tree-Partial view.

A decision tree is depicted in Figure 2.2 from a perspective that only shows portion of the tree. It gives a number of options to choose from, including the fact that one can keep playing if the outlook is good and the humidity is about typical.

It is common for a decision tree to be overfit when it is constructed using data from the training set. This indicates that the tree does an excellent

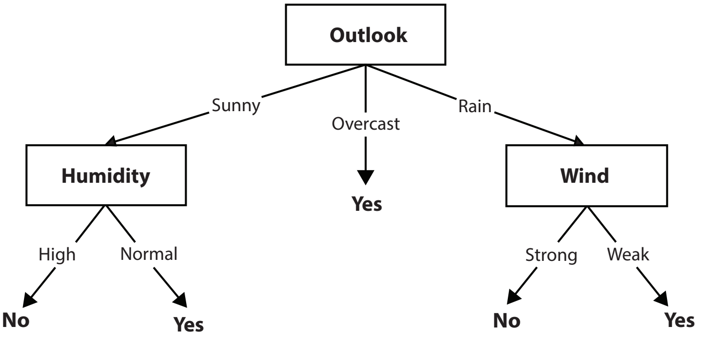


Figure 2.2 A sample decision tree—partial view.

job of training, as its name suggests. If the display contains content that isn't immediately apparent, the quality of the display as a whole will suffer. Therefore, it is possible to "shave off" hubs and sections of a decision tree, essentially substituting an entire sub tree with a leaf hub, if it comes to be established that the average error rate in the sub tree is greater than that in a single leaf. This is because "shaving off" refers to the process of removing a hub or section of a decision tree. The procedure of categorization is made easier as a result.

The process of performing trimming on a complicated decision tree aims to make it more controllable while also progressively expanding its scope. During a procedure known as "post trimming," some of the tree's branches are pruned away after it has reached its complete maturity, and younger trees are estimated with the help of test data. At this point, the ultimate tree will be chosen to be the one that is the least complicated and contains no mistakes.

Pre-pruning is another technique, which involves stopping the development of the tree at an earlier stage than normal. If a tree's overall soundness proportion is lower than a certain essential number, then the tree does not have a center. Choosing the appropriate accent, on the other hand, can be a difficult task. The C4.5 decision tree framework developed by Quinlan is a good example. The overwhelming majority of research involving decision trees focus on techniques to improve the presentation, such as clipping. In this article, we detail how we exploited data for insights into communities that have received inadequate research attention. Calculations based on decision trees are used in this situation to establish both graduation projections and the fundamental variables that contribute to graduation.

Included is a diagram illustrating the relationships between information extraction, the sharing of databases, and other topics such as artificial intelligence, measurements, and the visualization of data. It determines which of a large number of potential factors affecting a result are the most important, and then makes suggestions for strategies that can be used to organize demonstrations or forecast outcomes. It also describes in detail how the information that can be stored in a clustering framework's disorder network, which keeps track of both actual and intended configurations that can be used to evaluate the performance of a model.

In the context of a standard business task, information is presented in the shape of a table of tests, also known as instances. Each instance has a unique collection of distinguishing characteristics, also known as traits, as well as a designation that indicates the division to which it belongs. It describes an innovative method for evaluating characteristics by making use of genomic calculations as the primary tool.

The process of information extraction makes it possible to obtain data that has already been muddled and has the potential to be helpful. It is a form of instruction that makes use of computer algorithms to carry out an exhaustive search for recurring patterns in data. The extraction of useful knowledge from large amounts of data is one implementation of AI. Acquisition by association, or clustering, is an assumption of concept acquisition. The purpose of concept learning is to enable students to acquire an understanding of the significance of a general category after being presented with numerous instances, both positive and negative, of the category in question. Therefore, it draws a Boolean-valued competence conclusion based on the training occurrences. Unlike other methods of concept acquisition, arrangement learning can be expanded beyond the confines of a single educational setting. Finding models that are able to classify occurrences into the various preparatory instances that correspond to them is ultimately the objective of the learning process [16].

2.4 Machine Learning for Frequent Pattern Mining

To improve its effectiveness, the PSO algorithm will take use of the fact that many different kinds of animals, including fish and birds, exhibit behavior similar to that of a swarm. Within the parameters of the search, each particle has a position and velocity and these are completely unique in themselves, and they are free to move in any way that they see fit. The velocity and mobility of the particle are still constrained as a direct consequence of this, and the particle is directed to the location of any previously successful particles [17]. In addition to this, the particle is directed to the location of any other particles that have successfully completed their tasks in the past without causing any mishaps. It is now possible, as a result of the application of a predetermined set of rules to the technique in question, to characterize both the rate of motion and the location of each particle at every particular instant in time.

Identification 3 (ID3) is a method that use a decision tree to categorize instances in an iterative manner according to the values and features of those instances. It is generated from the most complex to the simplest using a collection of instances and the required features in that order. These groups were then further subdivided. When the set included inside a certain subtree is equal to the set of instances that belongs to a comparable category, the set in question is designated as a leaf node within the given subtree. Information theoretical criteria are taken into consideration whenever a characteristic is being chosen for testing purposes. The reduction in entropy will result in an increase in the information gain at each node [18].

Decision trees are often used in machine learning models for the purpose of data classification because of their speed and accuracy. When using this method, tree trimming may be accomplished in a number of different ways, making it a very adaptable practice. The result of the trimming is one that can be comprehended with ease. A number of academicians are of the opinion that taking down trees might be utilized as a means of overfitting. The C4.5 technique [20–22] includes an iterative classification that is used to further improve the categorization of the data until only pure leaf nodes are left, at which point the process comes to a conclusion. This iterative classification can be found in the method's outline. By using this strategy, it is possible to get the most out of the training data, while also eliminating the demand for rules [23, 24] that only signal a specific behavior in the data [19]. This is because it is practical to get the most out of the training data.

In recent years, the use of the technique known as the support vector machine (SVM) has seen significant growth in this sector. A helpful strategy for data categorization [25–27] is the use of a linear or non-linear separation surface in the input space. This may be done either linearly or non-linearly. The support vector machine (SVM) seeks to construct a model that reliably predicts the goal values of new data by using just the characteristics of the test data. The data collected during the training phase are used to construct the model. The support vector machine (SVM) provides a dependable approach for pattern categorization that maintains a high level of accuracy while simultaneously lowering the danger of overfitting. When there are just two types of classes present in the data, SVM may be used. Using SVM, data is categorized by locating the hyperplane that most effectively separates all of the data points into distinct categories. The hyperplane of a support vector machine (SVM) that has the greatest difference in margin between two classes of data is considered to be the optimal hyperplane for an SVM. When the slab is at its widest point parallel to the hyperplane, there are no data points located inside the boundary of the slab itself. Because of their close closeness to the hyperplane that separates the slab, the points along the slab's border are regarded to be support vectors. In fuzzy SVM, a membership value is assigned to each sample determined by its correlations with the values of other samples. There are many different ways in which each input sample contributes to the process that ultimately results in a choice being made. A radial basis function kernel has also been found to improve the performance of a support vector machine (SVM) [20].

2.5 Conclusion

Data mining is the act of applying complicated analytics to enormous datasets to discover hidden links, patterns, and trends. This has been done to uncover previously unknown information. The handling of huge amounts of data must be both efficient and adaptable to facilitate the retrieval of information and the making of well-informed decisions. It finds trends, correlations, or patterns in massive datasets that were not previously there. This is the objective of the process known as data mining. Since the inception of the World Wide Web, there has been a quick and large growth in the quantity of data that has been preserved and is available electronically. The capability of mining data for patterns and linkages that have not been identified before has become a vital resource in a variety of disciplines, including academia and business. As a result of the proliferation of the Internet, the quantity of data that can be kept and accessed online has increased at an exponential rate. As a consequence of this, methods for extracting insights from this mountain of data have become an essential component in a variety of disciplines, including educational institutions and commercial enterprises. It is standard practice to utilize a technique known as frequent item set mining to extract meaningful information from databases.

References

1. Haoxiang, W. and Smys, S., Big Data Analysis and Perturbation using Data Mining Algorithm. *J. Soft Comput. Paradigm (JSCP)*, 3, 01, 19–28, 2021.
2. Sivakami, M. and Prabhu, P., Classification of Algorithms Supported Factual Knowledge Recovery from Cardiac Data Set. *Int. J. Curr. Res. Rev.*, 13, 6, 161–166, 2021. ISSN: 2231-2196 (Print) ISSN: 0975-5241 (Online).
3. Bora, A., Gowri, NV., Naved, M., Pandey, P.S., An Utilization Of Robot For Irrigation Using Artificial Intelligence. *Int. J. Future Gener. Commun. Netw.*, 14, 1, 23–38, 2021.
4. Raghuvanshi, A., Singh, U., Sajja, G., Pallathadka, H., Asenso, E., Kamal, M. *et al.*, Intrusion Detection Using Machine Learning for Risk Mitigation in IoT-Enabled Smart Irrigation in Smart Farming. *J. Food Qual.*, 1, 2022, 1–8, 2022, doi: 10.1155/2022/3955514.
5. Jasti, V. *et al.*, Computational Technique Based on Machine Learning and Image Processing for Medical Image Analysis of Breast Cancer Diagnosis. *Secur. Commun. Netw.*, 2022, 1–7, 2022, Available: 10.1155/2022/1918379.
6. Hemamalini, V., Rajarajeswari, S., Nachiyappan, S., Sambath, M., Devi, T., Singh, B., Raghuvanshi, A., Food Quality Inspection and Grading Using

Efficient Image Segmentation and Machine Learning-Based System. *J. Food Qual.*, 2022, 1–6, 2022, doi: 10.1155/2022/5262294.

7. Raghuvanshi, A., Singh, U., Joshi, C., A Review of Various Security and Privacy Innovations for IoT Applications in Healthcare. *Adv. Healthcare Syst.*, 1, 43–58, 2022, Available: 10.1002/9781119769293.ch4.
8. Agrawal, R., Imielinski, T., Swami, A., Mining Association Rules between Sets of Items in Large Databases. *ACM Sigmoid Rec.*, 22, 2, 207–216, 1993, https://doi.org/10.1145/170036.170072.
9. Agrawal, R. and Srikant, R., Fast Algorithms for Mining Association Rules. *Proceedings of the 20th VLDB Conference Santiago*, Chile, pp. 487–499, 1994.
10. Al-bana, M.R. and Farhan, M.S., An Efficient Spark-Based Hybrid Frequent Itemset Mining. *Data (MDPI)*, 7, 11, 1–22, 2022, https://doi.org/https://doi.org/10.3390/data7010011.
11. Al-Maolegi, M. and Arkok, B., An Improved Apriori Algorithm For Association Rules. *Int. J. Nat. Lang. Comput.*, 3, 1, 21–29, 2014, https://doi.org/10.5121/ijnlc.2014.3103.
12. Magdy, M., Ghaleb, F.F.M., Mohamed, D.A.E.A., Zakaria, W., CC-IFIM: an efficient approach for incremental frequent itemset mining based on closed candidates. *J. Supercomput.*, 79, 7, 7877–7899, 2022, https://doi.org/10.1007/s11227-022-04976-5.
13. Ming-Syan, C., Jiawei, H., Philip, S.Y., Data Mining: An Overview from a Database Perspective. *IEEE Trans. Knowl. Data Eng.*, 8, 6, 866–883, 1996, https://doi.org/10.1109/69.553155.
14. Park, J.S., Chen, M.S., Yu, P.S., An Effective Hash-Based Algorithm for Mining Association Rules. *ACM Sigmoid Rec.*, 24, 2, 175–186, 1995, https://doi.org/10.1145/568271.223813.
15. Sandhu, P.S., Dhaliwal, D.S., Panda, S.N., Bisht, A., An improvement in apriori algorithm using profit and quantity. *2nd International Conference on Computer and Network Technology, ICCNT 2010*, pp. 3–7, 2010, https://doi.org/10.1109/ICCNT.2010.46.
16. Shuwen, L. and Jiyi, X., An Improved Apriori Algorithm Based on Matrix. *12th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*, pp. 488–491, 2020, https://doi.org/10.1109/ICMTMA50254.2020.00111.
17. Singh, H. and Dhir, R., A New Efficient Matrix Based Frequent Itemset Mining Algorithm with Tags. *Int. J. Future Comput. Commun.*, 2016, 355–358, 2013, https://doi.org/10.7763/ijfcc.2013.v2.184.
18. Sun, L.N., An improved apriori algorithm based on support weight matrix for data mining in transaction database. *J. Ambient Intell. Hum. Comput.*, 11, 2, 495–501, 2020, https://doi.org/10.1007/s12652-019-01222-4.
19. Mishra, A.K., Tyagi, A.K., Dananjayan, S., Rajavat, A., Rawat, H., Rawat, A., Revolutionizing Government Operations: The Impact of Artificial Intelligence in Public Administration, in: *Conversational Artificial Intelligence*, pp. 607–634, 2024.

20. Vivekanandan, S.J., Ammu, S.P., Sripriyadharshini, R., Preetha, T.R., Computation Of High Utility Itemsets By Using Range Of Utility Technique. *J. Univ. Shanghai Sci. Technol.*, 23, 4, 94–101, 2021.
21. Rathore, N. and Rajavat, A., Smart Farming Based on IOT-Edge Computing: Applying Machine Learning Models For Disease And Irrigation Water Requirement Prediction In Potato Crop Using Containerized Microservices, in: *Precision Agriculture for Sustainability*, pp. 399–424, Apple Academic Press, USA, 2024.
22. Patsariya, M. and Rajavat, A., A Progressive Design of MANET Security Protocol for Reliable and Secure Communication. *Int. J. Intell. Syst. Appl. Eng.*, *12*, 9s, 190–204, 2024.
23. Rathi, M. and Rajavat, A., Investigations and Design of Privacy-Preserving Data Mining Technique for Secure Data Publishing. *Int. J. Intell. Syst. Appl. Eng.*, *11*, 9s, 351–367, 2023.
24. Dubey, P. and Rajavat, A., Effective K-means clustering algorithm for efficient data mining. *2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN)*, IEEE, pp. 1–6, 2023, May.
25. Pithawa, D., Nahar, S., Bhardwaj, V., Rawat, R., Dronawat, R., Rawat, A., Quantum Computing Technological Design Along with Its Dark Side. *Quantum Comput. Cybersecur.*, 1, 295–312, 2023.
26. Shrivastava, M., Patil, R., Bhardwaj, V., Rawat, R., Telang, S., Rawat, A., Quantum Computing and Security Aspects of Attention-Based Visual Question Answering with Long Short-Term Memory. *Quantum Comput. Cybersecur.*, 1, 395–412, 2023.
27. Rawat, R., Chakrawarti, R.K., Raj, A.S.A., Mani, G., Chidambarathanu, K., Bhardwaj, R., Association rule learning for threat analysis using traffic analysis and packet filtering approach. *Int. J. Inf. Technol.*, *15*, 6, 3245–3255, 2023.

3

Classification and Detection of Prostate Cancer Using Machine Learning Techniques

D. Vetrithangam[1]*, Pramod Kumar[2], Shaik Munawar[3], Rituparna Biswas[4], Deependra Pandey[5] and Amar Choudhary[6]

[1]Department of Computer Science & Engineering University Institute of Engineering, Chandigarh University, Mohali, Punjab, India
[2]Computer Science and Engineering Ganga Institute of Technology and Management, Jhajjar, Haryana, India
[3]Department of CSE Kakatiya Institute of Technology and Science Warangal, Hanamkonda, Telangana, India
[4]Department of Basic Science and Humanities, IEM Newtown, UEM, Kolkata, India
[5]Department of ECE, Amity School of Engineering & Technology (ASET), Amity University, Lucknow, Uttar Pradesh, India
[6]Department of ECE, Alliance College of Engineering & Design (ACED), Alliance University, Bengaluru, Karnataka, India

Abstract

Carcinoma is a significant contributor to the death rates of individuals. Reducing the amount of time it takes to diagnose a patient is very necessary to improve their prognosis. Diagnostic imaging and other traditional methods are used by highly trained medical professionals to identify any telltale indicators that may be present in the bodies of their patients. In spite of the abundance of medical imaging data, manual diagnosis may still be subjective and time-consuming due to the fact that people's perceptions differ so much from one another. One of the primary reasons for the variability is the collecting of data from medical imaging. A proper diagnosis may be more difficult to get as a result of this. When performing activities such as machine learning and the processing of complex pictures, it is important to make use of the most advanced computational power available.

**Corresponding author*: vetrigold@gmail.com

Rajesh Kumar Chakrawarti, Ranjana Sikarwar, Sanjaya Kumar Sarangi, Samson Arun Raj Albert Raj, Shweta Gupta, Krishnan Sakthidasan Sankaran and Romil Rawat (eds.) Natural Language Processing for Software Engineering, (29–42) © 2025 Scrivener Publishing LLC

Ever since the 1980s, there has been a persistent effort to develop a computer-aided diagnostic system that has the potential to help in the early diagnosis of a wide variety of malignancies. According to the most recent estimates, around one-seventh of men will be diagnosed with prostate cancer at some point in their life. This illness claims the lives of so many men every year, and it is unbearable that the number of men who are diagnosed with prostate cancer continues to climb. It is a tragedy that this number continues to rise. A powerful diagnostic system that is capable of managing high-resolution, multi-dimensional MRI images is an absolute need, in addition to computer-aided design (CAD) software. In the present moment, I am focusing my attention on a project that will make it easier for us to achieve our shared goals. Scientists are now studying methods to improve the speed, accuracy, and precision of computer-aided design (CAD) technology since it has been shown to be valuable. CAD technology has been demonstrated to be effective, as shown by the evidence. The development of techniques for the diagnosis and classification of prostate cancer via the use of MRI image processing and machine learning is the fundamental objective of this study as well.

***Keywords*:** Prostate cancer, prediction, machine learning, feature extraction, image segmentation, accuracy

3.1 Introduction

The prostate, a relatively small organ that plays a crucial role in the male reproductive system, is an essential component of sexual reproduction. It is very important for the fluid that is generated by the prostate gland, which is referred to as semen, to assist in the passage of sperm through the male reproductive system. The most important thing is to make sure that it is situated between the bladder and the tube that is responsible for releasing pee from the body, which is the upper urethra. In spite of the fact that melanoma is more prevalent in males, prostate cancer (PC) is the main cause of death among men due to cancer. This is a significant reason for worry for the public health of the whole world. The unregulated expansion of cells inside the gland itself is the primary factor that leads to the development of prostate cancer [1]. There is a possibility that the pace of development of cancer in the peritoneal cavity will differ from patient to patient. The prostate is the only organ that can sustain the development of slow-growing tumors. It is believed that around 85 percent of all cases of pancreatic cancer are slow-progressing variations of the disease. It is very necessary to keep oneself in a state of heightened awareness to properly navigate these circumstances [2]. The second kind of pancreatic cancer has the potential to spread to other parts of the body, in contrast to the first type, which is less aggressive. It is only via the use of reliable monitoring methods [5] that it is possible to differentiate between these two types of development. Physical inspections performed on a regular basis may, in many instances, discover personal computers at an early enough level. It is essential to ascertain the precise anatomical location of the prostate gland to begin the process of formulating a treatment plan once the prostate gland has been identified. Only via the implementation of screening procedures that are dependable and efficient is it possible to obtain a high chance of survival. The utilization of magnetic resonance imaging (MRI), transrectal ultrasonography (TSUS), and prostate specific antigen (PSA) tests are often used in the process of screening for prostate cancer [3]. The provision of standard parameters for magnetic resonance imaging (MRI) was the major objective of the modifications that were made to the initial guidelines for prostate MRI. Both of these things were done in addition to concentrating on the categorization of clinical relevance. In accordance with the first directive, this is not something that should be done. With each new issue, the objective is to provide a higher standard of excellence for photography and journalism. There have been a number of studies conducted in recent times that have focused on the outcomes of notions that are founded on these principles. There are a few restrictions that need to be carefully addressed when diagnosing extremely tiny lesions that have incurred severe damage. However, it is feasible to categorize a PC lesion as clinically significant by utilizing one of these approaches. The findings of this research have significant implications for the staging of cancer since they demonstrate that the use of a PI-RADS technique may make it simpler to detect cancer that has spread beyond the prostate. This has occurred as a result of the infection spreading beyond the prostate, which is the reason for this [4]. The data that are maintained in biological databases, which are accessible to scientists, are a veritable treasure trove [15]. It is becoming more difficult to get actionable insights due to the ever-increasing volumes of data. The term "machine learning" refers to the process by which computers may enhance their intelligence and performance by watching and replicating human activities, comparing their findings to those of other datasets, and generating new information from start. The growth of this kind of learning was brought about by the widespread use of data mining, which is an essential component of knowledge mining. One of the most important aspects of machine learning is the ability to recognize patterns and quickly derive inferences from a wide variety of datasets. The evaluation of ligand libraries may be carried out automatically via the use of machine learning methods [6, 7].

3.2 Literature Survey

Rampun *et al.* [8] utilized a median filter and an anisotropic diffusion filter to achieve their results. Due to the similarity between noise and edge gradients in photographs with a low signal-to-noise ratio, removing noise can be a challenging task. Consequently, removing background noise from pictures becomes a challenging task. While it is possible to detect noise gradients using a thresholding approach, their limits are not as apparent as one might anticipate. Samarasinghe and colleagues [9] employed a three-dimensional sliding Gaussian filter in their study. A plethora of innovative and thorough alternatives to this scanning process have been proposed. These options might assist in addressing issues such as noise distribution in MPMRI images. Shrinkage methods can exploit the sparsity provided by wavelet decomposition to improve MPMRI images. Orthogonal transformations are utilized in the wavelet transform for practical purposes in the real world. This metamorphosis is occurring right in front of you. The unwanted noise signal might still be present when the Rician distribution is used in the wavelet transform. To account for the scattering of noise in the data, modifications were made to the wavelet and scaling parameters. Lopes *et al.* [10] employed a methodology that integrates detection and estimating techniques to remove noise from T2W pictures. The computation starts with the wavelet coefficients that contain noise and then proceeds with the wavelet coefficient that is free from noise, achieved using a posteriori maximum estimate. By resizing the pictures, it was guaranteed that the PZ region would have a standard deviation of zero after the normalization process. Subsequently, the instructional and evaluative uses of the normalized multi-parametric MRI images were examined. By employing defined dynamic ranges for the intensities of the MPMRI sequences, the authors were able to ensure stability in the segmentation process. We successfully achieved our goal by employing this approach. The MPMRI images seem distorted due to the presence of background noise and the bias field of the endorectal coil [11]. The presence of a bias field in MRI images may account for the observed variations in signal intensity. Consequently, there is a substantial disparity in the luminosity of similar tissues across different areas of the picture. As the computer-aided design process progresses, it becomes more and more demanding. Training images are essential as segmentation and classification are fundamentally educational processes. Image-based training is crucial. To develop a dependable automated diagnosis, it is crucial to gather photographs of individuals from the same group with similar signal intensities. Whether or not the patients have cancer is inconsequential. Variations in outcomes can occur across patients, despite the use of the same tools, protocols, and circumstances. Viswanath *et al.* [12] employed the piecewise linear normalization approach to standardize the T2W pictures. This enabled us to ensure that the outcomes would be uniform while also eliminating individual variances among patients. The aim of this approach was to get a homogeneous visual aspect in the photographs. One component of the investigation included identifying and replicating the original foreground by employing methods for piecewise linear normalization. Most medical image analysis algorithms heavily depend on atlases for the process of segmentation. Firstly, it excels in low-definition and pixel-resolution situations. Tian *et al.* [13] obtained the required results from their study of prostate MRI data by integrating the graph cut segmentation approach with the concept of superpixels. Utilizing copy-and-paste segmentation has the benefit of diminishing the requirements for memory and processing resources. Human interaction is necessary during setup as the process is only partially automated. Martin *et al.* [14] propose a method for reconstructing the prostate from MRI data using a segmentation technique that involves deformable models and atlases. The utilization of a malleable model enabled the atlas-based method to precisely align the shape of the prostate cancer with its surrounding boundaries. A likelihood-based strategy may be used to accurately detect the exact position of the prostate. The work conducted by Vincent and his colleagues [15] provides a comprehensive description of an automated method for segmenting the prostate in MRI data. The underlying idea was based on the concept of "active appearance". The optimization process employs a multi-start technique to enhance performance and ensure accurate alignment of the model with the test photographs. This is our approach to ensure that we meet all the requirements. The researchers utilized an atlas-based matching method to generate the autonomous partitioning of the prostate, as described by Klein *et al.* [16]. To do this, the target image was compared to an extensive collection of annotated atlas photographs. The investigation encompassed all aspects. Their registration procedure was lenient, and they employed manual segmentation and classification methodologies [28, 29]. By merging the corresponding pictures throughout the segmentation step, it becomes possible to achieve MR prostate image segmentation after the registration stage is completed. One potential strategy for prostate segmentation is to use deformable models that incorporate both internal and exterior energy sources. The internal energy of the prostate causes it to become flatter, while external energy helps to make it longer. The approach proposed by

Chandra *et al.* [17] for independently and effectively segmenting prostate pictures obtained by scanning does not require the use of an endorectal coil. This approach was devised with the aim of segregating the photos [30–32]. During the training phase of the initialization strategy, this case-specific deformable system generates a distinct triangulated surface and image feature system for each individual patient. Picture feature systems can alter the original appearance of an image by employing the technique of template matching. Another developing trend in automated prostate segmentation systems is the use of approaches that combine several atlases with deformable models. Yin and colleagues [18] employed a very dependable and fully automated technique for prostate segmentation in their experiment. The prostate mean shape system is improved by employing the graph-search approach. Upon employing a standardized gradient field to perform cross-correlation on the prostate, the procedure is concluded. The material presented here is expected to enhance our comprehension of the historical background and development of the prostate. Deformable models are beneficial in situations when noise or abnormalities in the data result in undesired borders of the prostate. By employing a graph-cutting methodology, one may obtain a thorough and unproblematic result without encountering any difficulties arising from excessive segmentation. The graph cut technique was initially employed for prostate segmentation, as stated in the work conducted by Mahapatra and colleagues [19]. The acquired semantic data is employed for this objective. Precise determination of the volume and location of the prostate can be achieved by employing a super-voxel segmentation method that relies on random forests. The use of images and environmental data to educate random forest classifiers facilitated the further enhancement of prostate volume. Utilizing a Markov random field for graph cuts in prostate segmentation has the potential to enhance their performance. Puech *et al.* [20] created a set of rules based on data obtained from hospital IT systems to forecast the outcomes of tests. The fundamental components of supervised learning for data organization are similarity metrics and k-nearest neighbor (k-NN) algorithms. K-means clustering is an unsupervised method for classifying datasets that involves dividing the data into k distinct groups through an iterative process. The variable "k" is used to express the number of iterations in this situation. The number of adjacent centroids defines the unique designation assigned to each point in the feature space. Once we have confirmed that the centroid coordinates of each cluster are the same, we next proceed to update the means and generate a new mean for each cluster. Once all possible modifications to the centroids have been taken into consideration, the process of assigning and adjusting them will ultimately lead to an accurate outcome. K is a renowned and often utilized notation for denoting the number of cluster classes. Linear discriminant analysis (LDA) is the most successful approach for separating the two groups, making it the preferred choice. As a result, disputes across different categories increase, but differences within the same category decrease. The Naive Bayes classifier is widely regarded as the most superior option by a significant number of individuals. Using the idea of independent feature dimensions, this classification method might be referred to as "probabilistic classification." By utilizing this strategy, it is possible to attain the highest achievable posterior probability for picture categorization. An frequently used method for classification is adaptive boosting, also known as AdaBoost. The AdaBoost ensemble learning algorithm has not been publicly released until [21]. The objective of this strategy is to create a robust classifier by merging many ineffective learners. AdaBoost (AdB) generally outperforms random forest (RF) when comparing the outcomes. This specific classifier has a preference for learning procedures that are not efficient, such as decision stumps, classification trees, and regression trees. Lopes and his colleagues employed an AdaBoost classifier to carry out the classification strategy for their inquiry. Class labeling, a type of sparse kernel-based classification, largely relies on Gaussian processes. This method is commonly known as the "kernel approach" since it utilizes the whole training dataset to generate new labels. Sparse kernel classification algorithms have the ability to accurately assign a label to an unknown picture by using only a tiny portion of the annotated examples in the training dataset [22]. A support vector machine (SVM), which is a sparse kernel approach, is used to identify the optimal linear hyperplane for separating data into two label classes and maximizing the margin of error. To accomplish this, classes were partitioned based on the most efficient linear hyperplane. Support vector machines excel as classifiers in real-world scenarios due to their robustness and ability to generalize.

3.3 Machine Learning for Prostate Cancer Classification and Detection

Here we will go over the machine learning methods that are employed in MR scans to identify prostate cancer. Image preprocessing makes use of the histogram equalization approach. You can see things more clearly now. For effective picture segmentation, the fuzzy C-means method is employed. Interactions at the Grey Level among the many possible approaches to feature extraction, matrix is one to consider. There are three classification approaches that get the most attention: K-nearest neighbors, Random Forest and Adaboost. Digital X-rays, MRI, CT, and PET images are just a few examples of the many imaging modalities that could benefit from histogram equalization, a basic methodology in image processing. When one uses this technique, the pictures get clearer and more detailed. It is critical to obtain high-quality photos to correctly diagnose the illness and use these images for diagnosis. If any noises were previously muted in the image, they may be brought to light during the histogram equalization step, which follows processing. This technique is commonly employed in medical imaging analysis, as stated in [23]. This method produces an imperfectly distributed histogram of grey levels by employing grey operations. To do this, the image's grey mapping is determined. Clustering is a method that may be used to find hidden correlations in photos by merging similar patterns into larger groups. Clustering is the process of organizing items into groups based on the shared qualities between them. The FCM technique generates membership values, which are utilized for data item classification. Next, the data are partitioned after optimizing the object function using the least squares technique [24]. Feature extraction is a method in image processing that might reduce the amount of data that has to be kept. To do this, we take a set of feature subsets and eliminate any dimensions that aren't relevant or needed. To restore the texture's characteristics without compromising the pixel-to-pixel correspondence, GLCM is the method of choice. One way to achieve this aim is to find the values of the co-occurrence of grey levels. When building the general linear model (GLM), one can choose a direction of ș = 0, 45, 90, or 135 degrees and a length d from 1 to 5. Generating the GLM is subsequently accomplished using the conditional probability density functions p(I|j, d, ş). To achieve this objective, the GLCM algorithm is employed. The physical distance between two samples, measured as the inter-sample distance (d), is one such example. It is also possible to find the likelihood that these samples are physically connected (i,j | d,ş) using a function that is another example. [25] Among the many noteworthy characteristics of the GLCM are its energy, homogeneity, correlation, and contrast.

One of the many areas where KNN excels is in classification issues; this is because it is a supervised approach. Remember that this method consistently yields the same results with each given set of training data. Using the value in the population that is most comparable to any of the samples, one may give a class to any or all of the samples. If one wants to know how close two pixel coordinates are to each other, he/she may use the following equation to find their Euclidean distance. All of these considerations lead to the same conclusion: grouping the pixels together from the start would have been better. Using the K-nearest neighbors (KNN) algorithm, one may find the neighborhood where the distance between any two neighbours is the lowest. Within the algorithm, this neighborhood is represented by the letter K. The number of nearby residences should be your first consideration. It is usual practice to have an even number of courses when there are just two. Specifically, at that stage of the process, the neighborhood nearest to the subject is calculated using the value K = 1. A consequence of this sort would be the most basic and uncomplicated kind [26].

The phrase "random forest" was created because the model may generate forests that appear to be random. Rarely will you hear RF called "random forest." Using this method, you can build a set of decision trees, and then you can teach each tree independently. All of the possible answers to the questions and many of the available selection possibilities are contained in the present "forest of trees," which was generated using this approach. This is why more precise estimates were produced by include their input into the computations [27].

For classifiers that aren't performing up to pace when it comes to data classification, an approach called "AdaBoost" can be employed to boost their performance. To begin, we will assign relative importance to each observation using the AdaBoost method. The subsequent iterations will place a greater emphasis on the correctly identified observations and a lesser emphasis on the erroneously classified ones. The results will prove this by showing that a clear trend exists. When it comes to increasing the effectiveness of the classifier, all that is necessary is the usage of weights, which reflect the class membership of an observation. As a consequence of this, mistakes in classification will reduce. With the "boosting" approach, which entails providing children with a series of tailored fits, it is likely that pupils who are experiencing problems in school could benefit from the treatment. The successive models [28] in the series add increased weight to the data that were previously rejected during the first phase of the series.

3.4 Conclusion

One of the leading causes of mortality among those aged 65 and older is cancer. There is a significant correlation between the date of a patient's diagnosis and the likelihood that they will survive. In addition to doing standard diagnostics, the trained eyes that evaluate medical pictures for indicators of cancer also perform this examination. These professionals are on the lookout for symptoms that might show that cancer is developing inside the body. Due to the vast quantity of medical imaging data and the wide range of inter-observer variability, manual diagnosis is a time-consuming process that is also susceptible to subjectivity. It is possible that the diversity might be linked to the substantial amount of data that is contained in medical imaging. It is more challenging to arrive at an accurate diagnosis as a result of this consideration. Image processing and machine learning operations that are very complex need the most advanced computer hardware available. Over the course of many decades, researchers have been working on a computer-aided diagnostic system with the intention of assisting medical professionals in the early identification of cancer. A diagnosis of prostate cancer is something that about one out of every seven men will get at some time in their lives. There has been an increase in the number of men who have been diagnosed with prostate cancer, and there has also been an increase in the number of deaths that have been caused by the illness. For the purpose of collecting high-fidelity and multi-faceted magnetic resonance imaging (MRI) images, it is necessary to have a diagnostic system that is suitable in conjunction with computer-aided design (CAD) capabilities. One of the projects that the researcher is now working on is designed to assist us in achieving our goals. Researchers are now concentrating their efforts on enhancing the speed, accuracy, and precision of the computer-aided design (CAD) technologies that are already available. With the use of machine learning, this study presents a model that is capable of accurately evaluating photographs, recognizing distinctive characteristics, and acquiring new abilities.

References

1. Vilanova, J.C., Catalá, V., Algaba, F., Laucirica, O. (Eds.), *Atlas of Multiparametric Prostate MRI: With PI-RADS Approach and Anatomic-MRI-Pathological Correlation*, Springer, USA, 2017.
2. Cameron, A., Khalvati, F., Haider, M.A., Wong, A., MAPS: a quantitative radiomics approach for prostate cancer detection. *IEEE Trans. Biomed. Eng.*, 63, 6, 1145–1156, 2016.
3. Jasti, V., Zamani, A., Arumugam, K., Naved, M., Pallathadka, H., Sammy, F. *et al.*, Computational Technique Based on Machine Learning and Image Processing for Medical Image Analysis of Breast Cancer Diagnosis. *Secur. Commun. Netw.*, 2022, 1–7, 2022, doi: 10.1155/2022/1918379.
4. Giannini, V., Vignati, A., Mirasole, S., Mazzetti, S., Russo, F., Stasi, M., Regge D.2016. MR-T2-weighted signal intensity: a new imaging biomarker of prostate cancer aggressiveness. *Comput. Methods Biomech. Biomed. Eng.: Imaging Vis.*, 4, 3–4, 130–134, 2022.

5. Chaudhury, S., Krishna, A.N., Gupta, S., Sankaran, K.S., Khan, S., Sau, K., Raghuvanshi, A., Sammy, F., Effective Image Processing and Segmentation-Based Machine Learning Techniques for Diagnosis of Breast Cancer. *Comput. Math. Methods Med.*, 2022, 6 pages, 2022, https://doi.org/10.1155/2022/6841334.
6. Weinreb, J.C., Barentsz, J.O., Choyke, P.L., Cornud, F., Haider, M.A., Macura, K.J., Margolis, D., Schnall, M.D., Shtern, F., Tempany, C.M., Thoeny, H.C., Verma, S., PI-RADS Prostate Imaging– Reporting and Data System: 2015, Version 2. *Eur. Urol.*, 69, 1, 16–40, 2016.
7. Abu Sarwar Zamani, L., Anand, K.P.R., Prabhu, P., Buttar, A.M., Pallathadka, H., Raghuvanshi, A., Dugbakie, B.N., Performance of Machine Learning and Image Processing in Plant Leaf Disease Detection. *J. Food Qual.*, 2022, 7 pages, 2022, Article ID 1598796, https://doi.org/10.1155/2022/1598796.
8. Rampun, A., Zheng, L., Malcolm, P., Tiddeman, B., Zwiggelaar, R., Computer aided detection of prostate cancer in t2-weighted mri within the peripheral zone. *Phys. Med. Biol.*, 61, 13, 4796–4825, 2016.
9. Samarasinghe, G., Sowmya, A., Moses, D.A., Semi-quantitative analysis of prostate perfusion mri by clustering of pre and post contrast enhancement phases, in: *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium*, pp. 943–947, IEEE, 2016.
10. Lopes, R., Ayache, A., Makni, N., Puech, P., Villers, A., Betrouni, N., Prostate cancer characterization on MR images using fractal features. *Med. Phys.*, 38, 1, 83–95, Jan 2011.
11. Styner, M., Brechbuhler, C., Szckely, G., Gerig, G., Parametric estimate of intensity inhomogeneities applied to MRI. Medical Imaging. *IEEE Trans.*, 19, 3, 153–165, 2000, ISSN 0278-0062.
12. Viswanath, S.E., Bloch, N.B., Chappelow, J.C., Toth, R., Rofsky, N.M., Genega, E.M., Lenkinski, R.E., Madabhushi, A., Central gland and peripheral zone prostate tumors have signi_cantly di_erent quantitative imaging signatures on 3 Tesla endorectal, *in vivo* T2-weighted MR imagery. *J. Magn. Reson. Imaging*, 36, 1, 213–224, Jul 2012.
13. Tian, Z., Liu, L., Zhang, Z., Fei, B., Superpixel-based segmentation for 3D prostate MR images. *IEEE Trans. Med. Imaging*, 35, 3, 791–801, 2016.
14. Martin, S., Troccaz, J., Daanen, V., Automated segmentation of the prostate in 3D MR images using a probabilistic atlas and a spatially constrained deformable model. *Med. Phys.*, 37, 4, 1579–1590, 2010.
15. Vincent, G., Guillard, G., Bowes, M., Fully automatic segmentation of the prostate using active appearance models, in: *Proceedings of the 15th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) Grand Challenge: Prostate MR Image Segmentation 2012*, p. 7. Nice, France, 1–5 October 2012.
16. Klein, S., Van Der Heide, U.A., Lips, I.M., Van Vulpen, M., Staring, M., JP, P., Automatic segmentation of the prostate in 3D MR images by atlas matching using localized mutual information. *Med. Phys.*, 35, 4, 1407–1417, 2008.

17. Chandra, S.S., Dowling, J.A., Shen, K.-K. *et al.*, Patient specific prostate segmentation in 3-Dmagnetic resonance images. *IEEE Trans. Med. Imaging*, 31, 10, 1955–1964, 2012.
18. Yin, Y., Fotin, S.V., Periaswamy, S., Kunz, J., Haldankar, H., Muradyan, N., Cornud, F., Turkbey, B., Choyke, P., Fully automated prostate segmentation in 3D MR based on normalized gradient fields cross-correlation initialization and LOGISMOS refinement, in: *Medical Imaging 2012: Image Processing*, vol. 8314, p. 831406, International Society for Optics and Photonics, USA, 2012.
19. Mahapatra, D. and Buhmann, J.M., Prostate MRI segmentation using learned semantic knowledge and graph cuts. *IEEE Trans. Biomed. Eng.*, 61, 3, 756–764, 2014.
20. Puech, P., Betrouni, N., Makni, N., Dewalle, A.S., Villers, A., Lemaitre, L., Computer-assisted diagnosis of prostate cancer using DCE-MRI data: design, implementation and preliminary results. *Int. J. Comput. Assist. Radiol. Surg.*, 4, 1, 1–10, Jan 2009.
21. Freund, Y. and Schapire, R., A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55119–139, 1, 68–78, 1997.
22. Bishop, C.M., *Pattern recognition and machine learning*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.
23. Kalhor, M., Kajouei, A., Hamidi, F., Asem, M.M., Assessment of Histogram-Based Medical Image Contrast Enhancement Techniques; An Implementation. *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 0997–1003, 2019, doi: 10.1109/CCWC.2019.8666468.
24. Vela-Rincón, V.V., Mújica-Vargas, D., Mejía Lavalle, M., Magadán Salazar, A., Spatial αα-Trimmed Fuzzy C-Means Algorithm to Image Segmentation, in: *Pattern Recognition. MCPR 2020. Lecture Notes in Computer Science*, vol. 12088, K. Figueroa Mora, J. Anzurez Marín, J. Cerda, J. Carrasco-Ochoa, J. Martínez-Trinidad, J. Olvera-López (Eds.), Springer, Cham, USA, 2020, https://doi.org/10.1007/978-3-030-49076-8_12.
25. Benco, M., Kamencay, P., Radilova, M., Hudec, R., Sinko, M., The Comparison of Color Texture Features Extraction based on 1D GLCM with Deep Learning Methods. *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pp. 285–289, 2020, doi: 10.1109/IWSSIP48289.2020.9145263.
26. Uddin, S., Haque, I., Lu, H. *et al.*, Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. *Sci. Rep.*, 12, 6256, 2022, https://doi.org/10.1038/s41598-022-10358-x.
27. Jackins, V., Vimal, S., Kaliappan, M. *et al.*, AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. *J. Supercomput.*, 77, 5198–5219, 2021. https://doi.org/10.1007/s11227-020-03481-x.

28. Mahesh, T.R., Dhilip Kumar, V., Vinoth Kumar, V., Asghar, J., Geman, O., Arulkumaran, G., Arun, N., AdaBoost Ensemble Methods Using K-Fold Cross Validation for Survivability with the Early Detection of Heart Disease. *Comput. Intell. Neurosci.*, 2022, 11 pages, 2022, Article ID 9005278 https://doi.org/10.1155/2022/9005278.
29. Rathi, M. and Rajavat, A., Investigations and Design of Privacy-Preserving Data Mining Technique for Secure Data Publishing. *Int. J. Intell. Syst. Appl. Eng.*, *11*, 9s, 351–367, 2023.
30. Dubey, P. and Rajavat, A., Effective K-means clustering algorithm for efficient data mining, in: *2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN)*, IEEE, pp. 1–6, 2023, May.
31. Chauhan, D., Singh, C., Rawat, R., Chouhan, M., Conversational AI Applications in Ed-Tech Industry: An Analysis of Its Impact and Potential in Education. *Conversational Artif. Intell.*, 1, 411–433, 2024.
32. Chauhan, D., Singh, C., Rawat, R., Dhawan, M., Evaluating the Performance of Conversational AI Tools: A Comparative Analysis. *Conversational Artif. Intell.*, 1, 385–409, 2024.

4

NLP-Based Spellchecker and Grammar Checker for Indic Languages

Brijesh Kumar Y. Panchal[1,2*] **and Apurva Shah**[3]

[1]Computer Science and Engineering Department, Faculty of Technology and Engineering, The Maharaja Sayajirao University of Baroda, Vadodara, Gujarat, India
[2]Computer Engineering Department, Sardar Vallabhbhai Patel Institute of Technology (SVIT)-Vasad, Gujarat Technological University (GTU), Anand, Gujarat, India
[3]Computer Science and Engineering Department, Faculty of Technology and Engineering, The Maharaja Sayajirao University of Baroda, Vadodara, Gujarat, India

Abstract

Natural language processing (NLP) is a field that combines linguistics with computation. It falls under the umbrella of artificial intelligence (AI) and has prompted the development of a related topic known as computing linguistics. This field focuses on natural language processing in computing systems. A number of sentences, which are linguistic units consisting of one or more words connected based on a predetermined set of rules known as grammar, comprise a natural language. Grammar checking refers to the process of verifying the syntactic correctness of phrases and is a widely used technique in the field of language engineering. Spellchecker is an application that examines potential instances of incorrect spelling within a given text. It identifies and occasionally recommends keywords spelled incorrectly in a given text. The rate of mistake detection and error correction increases as the size of the spellchecker's lexicon increases. The goal of this study is to look at the techniques and procedures used by spellcheckers and grammar checkers for Indian languages, with the ultimate goal of producing comprehensive literature. Thus, one can create a more accurate system. This study

**Corresponding author*: panchalbrijesh02@gmail.com

Rajesh Kumar Chakrawarti, Ranjana Sikarwar, Sanjaya Kumar Sarangi, Samson Arun Raj Albert Raj, Shweta Gupta, Krishnan Sakthidasan Sankaran and Romil Rawat (eds.) Natural Language Processing for Software Engineering, (43–70) © 2025 Scrivener Publishing LLC

concludes by examining the many characteristics present in current spellcheckers and grammar checkers targeting Indian language.

***Keywords*:** Spellchecker, grammar checker, Indic languages, natural language processing (NLP), rules-based, machine learning, artificial intelligence (AI)

4.1 Introduction

Individuals often make grammatical as well as spelling errors when composing articles. This issue mostly affects those who are not good at the rules of language writing, leading to a significant investment in work and time in identifying and correcting spelling and grammatical errors. Therefore, individuals need a software application that facilitates automated grammatical and spelling checking while writing. Spellcheckers and grammar checkers are significant aspects of the study of natural language processing (NLP), and it is considered a well-established problem in NLP. The primary focus of this task is to rectify a range of faults in the text, including but not limited to spelling, punctuation, grammatical, and word-choice errors. This chapter introduces NLP-based spellchecker and grammar checker techniques in 1st part; the 2nd part describes a grammar checker scenario; and the 3rd part discusses the study of grammar checkers in the context of Indic languages.

4.2 NLP-Based Techniques of Spellcheckers and Grammar Checkers

4.2.1 Syntax-Based

Under this method, the provided text is parsed fully. Upon parsing, each phrase is given a parse tree depending on the grammar of the base language. If full parsing fails, the text is regarded as being wrong. Hence, the parser should be as thorough as possible in order to lower the number of false alarms. The key benefit of this strategy is that regardless of the kind of mistake, the grammar checker will recognize all wrong sentences if the grammar supplied is full, that is, if it covers all conceivable syntactic rules of the language. Unfortunately, it is impossible to clearly enumerate all of a natural language's syntactic rules owing to ambiguities in natural languages. Therefore, even for phrases that are accurate, the parser may yield more than one parse tree. This method only allows for the detection of false sentences. Nevertheless, additional rules that parse poorly constructed sentences are needed to inform the user of the issue; this method is referred to as constraint relaxation. A statement is considered wrong and a rule description and recommendation may be offered if it can only be parsed using this additional rule.

4.2.2 Statistics-Based

In this technique, a list of part-of-speech (POS) tags is provided using an annotated corpus. Certain sequences from these created sequences will be highly frequent, while others are likely to not appear at all. Sequences that often occur will be accepted as accurate in other texts, but rare sequences will result in mistakes. The existence of a sizable volume of POS-annotated data is a crucial prerequisite for using this technique. Also, the grammatical characteristics necessary for determining an agreement in the underlying language must be reflected in the employed POS tags.

4.2.3 Rule-Based

In this manner, a text that has at least been POS-tagged is compared against a set of predetermined rules in the form of mistake patterns. If one of such patterns is identified, the text is incorrect. Words, their POS tags, or even chunk tags may be used directly as the foundation for patterns. This scheme is similar to the statistics-based plan, except that all the rules are created by hand. Contrary to syntax-based systems, rule-based systems will never be finished. Even if there are many mistake rules present, it is almost impossible to anticipate every grammatical inconsistency; therefore, there will always be some errors that it misses. It is nevertheless preferable to have some problems go undiscovered than to have an insufficient parser generate unwanted false alarms. This method differs from others in that each rule may be individually enabled or disabled, and the system can provide detailed error messages along with beneficial comments, including explanations of grammatical rules. Using this strategy, the system may be expanded gradually by beginning with only one rule and adding more rules one at a time.

4.2.4 Deep Learning-Based

Artificial neural networks, which are machine learning (ML) algorithms, specialize in deep learning. Deep learning methods have recently been reported to be extensively used and to have achieved effective outcomes. One of the key factors in the achievement of deep learning methods is the freedom they provide when selecting the architecture. In ML research for natural language processing, deep learning approaches were at the forefront. Deep learning, a relatively new topic, goes beyond symbols to represent words as vectors. The semantic meaning of words is encoded in vectors. Deep learning learns vectors for words. One may quantify the distance between the vectors for two words rather than merely assuming that they are equal or unequal, as with symbols, and this measure of distance can provide a potent mechanism for generalization. Deep learning may discover, for instance, that the meaning of the phrase "I feel concerned" is more similar to that of "I feel nervous" than to "I feel tired." A grammar checker built on top of deep learning is called DeepGrammar. It uses deep learning to build a model of language, which it then applies in three phases to check text for errors: (1) Calculate the probability that the text was intentionally written. (2) Try to create wording that is more plausible and similar to the original material. (3) Show the user the text as a potential correction if such material is discovered.

4.2.5 Machine Learning-Based

Machine learning (ML) and NLP are key components in the course of transforming an artificial agent into an artificial "intelligent" agent. An artificially intelligent system can now process more information from its environment and reply to it in a manner that is easy for users to understand because of advancements in natural language processing. Natural language must pass through many stages of processing before a computer can understand it. This method consists of many steps, including morphological analysis, syntactic analysis, semantic analysis, discourse analysis, and pragmatic analysis. People often apply these analytical activities sequentially. In some ways, ML significantly improves almost all of these NLP procedures.

4.2.6 Reinforcement Learning-Based

Machine learning includes reinforcement learning, making it a subset of artificial intelligence. It enables software agents and machines to autonomously decide the best course of action within a certain situation to optimize performance. The reinforcement signal, sometimes referred to as easy reward criticism, is necessary for the agent to learn its behavior. As a subfield of ML and an extension of artificial intelligence, reinforcement learning makes it possible for machines and software agents to automatically decide what behavior is most appropriate in a given situation optimize performance. The reinforcement signal, which is straightforward reward feedback, is needed for the agent to learn its behavior.

4.3 Grammar Checker Related Work

In 2002, this study [1] discussed that the two-pass parsing method is essentially established to cut down on repetition in the grammar rules created for sentence analysis' phrase structure. The sentence is first parsed using some basic phrase structure grammar (PSG) rules. Failure results in the application of the movement restrictions and a reparative punishment. The approach examines declarative sentences for grammatical and structural errors. Implementing two-pass parsing is a novel and distinctive way to address challenging parsing issues. It enables you to maintain a simpler computational grammar while yet covering the widest possible variety of phrases. Without truly delving into the specifics of transformation grammar, it offers you a taste of how transformations work. When a declarative Urdu phrase is entered, the developed system will check it for grammar faults, and if any are discovered, it suggests fixes. The system's implementation validates the two-pass parsing strategy and the suggested learning algorithm.

In 2002, the researchers [2] presented the computational model that will be used to create a grammar checker for Urdu. The model employs the two-pass parsing technique to analyze sentences. The two-pass parsing method is primarily used to cut down the redundancy of the grammar rules for phrase structures that were developed for the analysis of sentences. At first, some base PSG rules are used to analyze the sentence. If the sentence cannot be parsed, movement rules are applied and then rewritten. The model is used to check for grammar and structural errors within declarative sentences. Two-pass parsing is a unique and innovative method for solving complex parsing challenges. It helps you make computational grammar simpler while covering a full range of sentences. Furthermore, it lets you experience a number of transformations and functions without getting into the transformational grammar's specifics. Finally, researchers have implemented the model. The system implemented can take an explicit Urdu sentence as an input and then check its grammar. If any errors are detected, it provides optional corrections to the incorrect sentence. This system is implemented to prove the two-pass parsing method and the algorithm proposed for the computational model.

In 2007, this study [3] developed the Nepali Grammar Checker, which is undergoing testing and development. It is described in this chapter along with its architectural and system design. This grammar checker was developed using a modular methodology and is made up of several components. The overall integrated system then uses these components as a pipeline. It is designed to look for grammatical problems in Nepali, including nominal and verbal agreements, parts of speech inflections, phrase and clause structures, and many types of language structures. The study and production of the first two modules—the tokenizer and the morphological analyzer—of the modules suggested in the architecture and system design of the Nepali Grammar Checker have been partly finished. The two modules have prototypes that are being tested including their complex management capabilities. As new information from their search effort becomes available in the future, the suggested architecture and system design of the Nepali Grammar Checker may vary.

In 2008, this article [4] described the grammar-checking system that was urbanized to detect grammatical mistakes within Punjabi texts. The system employs the full-form lexicon to analyze the morphology and uses rules-based techniques for POS tagging and phrase chunking. The scheme takes an innovative move toward an agreement to ensure that at the level of clause and phrase, the grammatical information provided through POS tags in the form of feature–value pairs is analyzed. The scheme can notice and advise corrections for various grammatical mistakes that result from the absence of an agreement or the arrangement of words in different phrases and phrases, etc. To our knowledge, this grammar-checking system is the first system to be reported in Indian languages. This chapter outlined the specifications for designing and operating the grammar-checking system for Punjabi. This system can detect numerous grammatical mistakes found in official Punjabi documents. As far as the researchers know, this is the first system of its kind applicable to Punjabi and different Indian dialects. The researchers hope this study will help close the gap in Punjabi and other languages in NLP. In addition, this research will inspire future researchers to continue creating advanced tools supporting Punjabi. This system aims to detect different grammatical errors in legal texts written in Punjabi. Identifying grammatical errors primarily focuses on ensuring that false alarms are kept to a minimum. The system gives enough details for each detected error so the user can understand what the error is and why it is being identified. The system also offers suggestions if it is possible to correct the errors.

This chapter outlined an innovative method of performing grammar checks with the help of clause and phrase information paired with grammatical information (POS data) in the form of feature value. This method can be used for languages with advanced resources, such as a full-parser pattern, and the pattern-matching methods cannot recognize other agreements. The online version of this grammar checking is accessible for free with three additional resources for the Punjabi language: the morphological analyzer, the tagger for POS, and the word chunker. The morphological analyzer is available as a free download for non-commercial usage.

In 2009, this study [5] focused on Punjabi language POS and grammar checker, an associate of the Modern Indo-Aryan family of languages, and is the subject of this study's POS tagging research. It was suggested to utilize a tagset for applications like grammar checking. This fine-grained tagset is completely based on the grammatical groups concerned in several types of concord in ordinary Punjabi sentences. The inflectional morphology of Punjabi words provide the basis for many of the morpho-syntactic traits used in this tagset. Because there is not a tagset for Punjabi or other Indian languages, this one was created with an emphasis on the agreement aspects of these languages. The grammar elements needed for agreement testing in Punjabi documents are not entirely covered by the tagsets for other languages. Also discussed is a rule-based tagger that was created using this tagset. This will be the first POS tagger for Punjabi to be released. For the languages that share grammatical traits with Punjabi, more especially the languages of the Modern Indo-Aryan family, the tagset given in this chapter is suggested for grammar testing and other purposes of a comparable kind.

In 2012 [6], researchers presented a unique method of "Hindi" grammar checker in their paper. The system used a complete dictionary for morphology analysis as well as rules-based systems. In this method, the researchers suggested a system that uses rules that is compared to the input Hindi sentence at a minimum and tagged. This method is similar to the approach based on statistics. However, all rules are designed by the researchers themselves. This rule-based Hindi Grammar Checker system was implemented productively for basic Hindi sentences only. The marks are quite positive. The benefit of this method is that the time it takes for grammar checks and analysis of the entire sentence is much less than other schemes. The reason for the higher performance is the fact that it will only look at patterns that are the same in terms of the words that are present in an input sentence, and does not take into account all patterns that are stored in the database. The researchers said that the system performance is further enhanced by adding more rules to the database manually.

In 2014, this paper's researcher [7] planned a spellchecker and grammar checker system designed to detect spelling errors and correct errors in text or a text using the hybrid approach. Grammar checker is an application that helps to determine grammar errors within the written text. Making a spellchecker or grammar checker program for Indian languages like Punjabi creates new problems that cannot be found in English, making the design of spellcheckers and grammar checkers a challenge. The biggest hurdles to overcome are that there is no fundamental layout for the Punjabi keyboard and no standard format approved for Punjabi spelling. There are numerous different grammatical features of Punjabi which makes it distinct when compared to other languages. Punjabi is the 12th most popular spoken language. The most important requirement for creating any spellchecker is to have a dictionary of various words from the language that can function as a lexicon. This paper sought to create an application that is a hybrid mix of grammar checking for Punjabi only. The system first checks for spelling errors and later for grammar errors within the content. If a text input is sent to the system, it goes through grammar and spell checkers before moving on to a grammar checker. A scalable, comparable algorithm was proposed that is a hybrid of grammar checker and spellchecker for Punjabi, which reduces time and costs. The purpose of the paper was to suggest an approach that is a mix of grammar checker and spellchecker for the Punjabi language. The projected system initially checks spelling errors, and after correcting spelling mistakes, it examines grammatical mistakes. The output is a document that is spelling error-free and grammar error-free. The correctness of the system is 83.5%. As per the researcher of this paper, this system can be improved to handle more complex sentences in the future. The system's accuracy was also improved. Other researchers could utilize this system for other languages, too.

In 2015, this study [8] provided a grammatical model analysis of Assamese sentences, which grew from general computational, linguistic, and psycholinguistic study. Linguists and NLP practitioners rely on parsing to decipher the syntactic and semantic details of a language's grammar. Free world order, ambiguity, and inefficiency are only some of the issues that make natural language text difficult to parse. For the Assamese language, the researcher suggested a model using the top-down parsing technique and a context-free grammar (CFG) with a restricted vocabulary. In this research, the researcher described a top-down predictive parsing method for analyzing basic Assamese phrases. Our suggested grammar for Assamese sentences only takes into account six different sorts of elements of speech. Future studies have to take into account as many possible sentence patterns in Assamese as well as other aspects of the language.

In 2015, the researcher [9] worked on developing Tamil morphological generators and analyzers. Making a comprehensive and effective operating system is quite different from making a morphological generator or analyzer for demonstration purposes. Making a system that provides 60% to 65% coverage is not too difficult. It is challenging to raise the efficiency over 65% and to a realistic level of 95% to 97%. Instead of being broad and detailed, the majority of the current descriptions concentrated on a vertical feature that is limited. The coverage that was attained included the descriptions of the previous work, which is often never more than 50%. The assessment technique revealed that the effectiveness of the morphological analyzer created by the current team for Tamil is really impressive. The creation of a morphological analyzer, which includes the analysis of verbal complex, has several natural language applications, including lemmatization, text production, machine translation, and parsing. Additionally, it aids document retrieval and word processing applications like spell checking and text input as well as voice applications like text-to-speech (TTS) synthesis and speech recognition. Assigning POS tags to the verbal complex is made easier by locating the categorical details of the verbal forms.

In 2016, this research paper [10] presented methods that can be utilized to enhance the current Punjabi Grammar Checker. In this study, two key actions, *i.e.*, the morphological analyzer and POS tagger, were enhanced to improve the efficiency of the current Punjabi Grammar Checker. After enhancing these two functions, development of significant magnitude in the current Punjabi Grammar Checker was noticed. In this study, a cross approach was adopted to ensure POS tag disambiguation in Punjabi. The researcher employed an amalgamation of rule-based and statistically-based methods in this method. The hybrid system was implemented with two distinct phases. The first output from the morph was provided as input to the rule-based POS tagger. In the case of a rule-based POS tagger, it was implemented in an exacting order. Every rule is paired with an initial condition known as an entry condition. If the first condition is satisfied, the rule will be relevant, and the researcher will follow the next rule. A common rule is changing tag A to B subject to the previous tag being C. This rule checks the first condition when the preceding tag is C. Then, relate this rule to modify tags from B to A or be relevant to the next rule. The first conditions for the applicability of rules usually depend on words, tags, and other morphological characteristics. For instance, the rules could be to change tag A to tag B with the first requirement being the preceding word contains the word "W." In this stage, the majority of the confusion can be resolved by the handwritten rule. In the next phase, the production of rule-based taggers for POS was used as input to the hidden Markov model (HMM) statistical tagger for POS. In this stage, all ambiguities that still need to be solved by the rule-based system are clarified. In the statistical method, a bi-gram hidden Markov model was employed. The annotated text of about 20,000 words was used to train and estimate the HMM parameter. An annotated corpus is built using a rule-based tagger. The method of maximum likelihood was employed to determine the parameter. The researcher used the Viterbi algorithm to create this method.

In 2016, research [11] techniques for enhancing the current Punjabi Grammar Checker were offered in this study work. To improve the presentation of the current Punjabi Grammar Checker, two primary activities—the morphological analyzer and the POS tagger—were upgraded in this study. Development and creation in the current Punjabi Grammar Checker were seen after upgrading these two tasks. First, the morphological analyzer and POS tagger, two upgraded modules, were examined separately. The grammar checker was then evaluated once the upgraded modules were swapped out for the standard ones. The morphological analyzer was much improved, as indicated in paper. The percentage of words that the morphological analyzer failed to identify decreased by 5–6%. The study also demonstrated an 8–9% development in the disambiguation. Despite an additional 0.5% to improper disambiguation, this may had been overlooked since disambiguation was much improved. Again, according to other study, the morphological analyzer and the POS tagger both contributed to a 27% and 20% decrease in false alarms, respectively.

In 2017, this paper's [12] author designed a grammar-checking system that works for the Hindi language. All of the elements needed for developing a grammar checker were designed from scratch. Certain mechanisms such as morph, POS tagger, and error detection systems were created using a statistical approach, and grammar correction systems were created using a rule-based approach. This system was tested against four types of mistakes. In an unnatural test of sequence-to-sequence strategies for sentence correcting, the researcher discovered that models based on characters tend to be more efficient than models based on words and models that encode sub-word information through convolutions, and modelling the results as a sequence of diffs increases the efficiency over conventional techniques. The most effective sequence-to-sequence model improves over our strongest phrase-based machine translation models, which has access to similar information in the range of six M2 (0.5 the GLEU) points. In addition, in the CoNLL-2014 standard setup's data-driven environment, the researcher showed that the modelling diffs can yield comparable or higher M2 scores when using simplified models or significantly smaller data sets than prior sequence-to-sequence methods. Our research shows that, on an extensive, professionally-annotated data set, a sequence-to-sequence model based on the character can result in significant efficiency advantages over a current smart machine text (SMT) system that includes task-specific features such as ceteris paribus. Additionally, when you consider the crowd-sourced environment of CoNLL data, where there are only a few sentences that have been professionally annotated in training, modelling diffs provide the possibility of making adjustments that increase the effectiveness of the sequence-to-sequence model for the job.

In 2018, in this paper's [13] review of rule-based Gujarati grammar the researchers said that NLP is the abbreviation. It has become a hot topic for study. The study of NLP and its applications reveals how a computer can comprehend natural language speech and text and modify it to perform amusing and practical tasks. Natural language processing (NLP) uses the word "language" to refer to natural languages like Gujarati, Hindi, and English. The main goal of this work is to evaluate the tech that is already in use and to determine the best way to use some of its NLP applications to construct Gujarati grammar. Our major emphasis is on the rule-based function since Gujarati has its own set of specific rules for mixing the consonants, vowels, and modifiers, much like every other language written in Indian script. The development of Gujarati grammar using current technology and appropriate NLP ideas, similar to other languages, is the primary emphasis of this study. The Gujarati language's grammar really requires a lot of labor to execute. The consistency of the Gujarati language and its grammar is the biggest practical issue. Here, the researcher attempted to at least provide a clear path for some of the NLP's language-based rules. Any researcher may in the future utilize a rule-based morphological strategy for classifying Gujarati and expand on this work by implementing grammar rules that may be utilized for other studies in the field of NLP.

In 2019, researchers of this paper [14] tested this system and it showed general accuracy of 0.83, recall of 0.91, and F-measure of 0.87. The grammar-checking software reviewed input sentences. If the sentence was not grammatically correct, then the program provided a suggestion for correcting the sentence if the developed system can check each sentence one at a time. The sentence was then ended with a sentence ending. The system developed was primarily created for sentences with a simple structure in the Hindi language, however, certain compound sentences can be processed to identify and correct mistakes. If the sentence was valid, it was displayed correct. In this research, the author suggested a hybrid method to create a grammar checker tool for identifying and correcting grammatical errors in the Hindi language. This can be described as the fundamental design for the Hindi Grammar Checker. The proposed system can detect and correct grammar mistakes caused by a lack of agreement between nouns and verbs in terms of numbers and gender. The system is designed to work only for short sentences in the Hindi language. The system also checks for issues relating to the lack of agreement between verb and noun or adjective and noun. Therefore, the researcher could further extend the research work to more complex sentences, *i.e.*, complex sentences and compound sentences in Hindi. Further, algorithms for detecting and correcting other kinds of errors, such as an error in the style, an error caused by postposition, and the use of words that are not necessary, could also be used.

In 2020, this paper's [15] researcher said that modern society's reliance on language as its primary means of communication is rooted in its long and storied past. The quality of a language depends in large part on its grammar. Researchers were schooled our whole lives to understand and communicate with one another via a body of information researchers have amassed, learned in accordance with rules, and a limited relevance. Furthermore, the ability to transfer this level of understanding into a computer, where it can interpret and categorize contextual evidence into the correct syntactical form, thereby validating that the information was in the correct form, is incredibly important right now due to the complexity of the task involved. This study discussed the problem and argued that the Dravidian language Kannada might benefit from the development of a tool to check its grammar. Among the first explanations would be the fact that the language's complexity presents a challenge, and therefore using a rule-based approach is a simpler way to go and makes it feasible to adequately identify recognized problems. A language expert is needed to construct hundreds of parallel standards that are hard to maintain. In this work, researchers proposed a model that makes use of a deep learning technique to train an LSTM (long short-term memory) neural network using a large dataset, with the help of Word2Vec, TensorFlow, and the Keras packages to retain information contextually. The suggested approach is efficient at detecting grammatical errors. In this study, researchers provided a deep-learning model for the Kannada language that uses the Word2Vec technique for granular embedding and was then trained in a neural network with an LSTM layer, which keeps the word's etymology in mind while learning its meaning. One shortcoming was the lack of a sizable annotated corpus for the Kannada language. The process of stemming and lemmatizing the data would further improve the paper by allowing for more uniform word interpretation. Additional work is needed to test the system on a much larger dataset in order to fine-tune the model and to make it easily available for future research.

In 2020, this study [16] worked toward constructing a flexible and comprehensive Arabic auditor that can deal with vowelized texts. An automatic grammar checking system may enhance the excellence of the text, lower the expenses of proofreading, and contribute to grammar instruction. The "Arabic Grammar Detector" was termed (AGD). Based on an addiction grammar and decision tree classifier model, AGD was effectively built. Its goal was to identify the proper syntactic dependencies of a phrase by extracting patterns of grammatical rules from a projective addiction network. The present implementation included practically all standard Arabic grammar rules for both vowelized and non-vowelized texts. Utilizing the Tashkeela corpus, AGD was assessed. More than 94% of grammatical problems were caught and fixes were also suggested. An Arabic grammar auditor is a complicated system that needs the assistance of experts in the area as well as substantial linguistic study and resources. This study presented an in-depth grammar auditor that uses a novel methodology. It seeks to identify grammatical mistakes in Arabic texts that include vowels and those that do not, as well as suggest how to fix them. This technique is based on the projective dependency graph's grammar, whose rules are taken from the hierarchy of grammatical rules that authors have created. In essence, the dependencies are synthesized in order to identify the most precise pattern of grammatical rules based on the properties of the sentence words and to determine the correct end-mark based on the output of the decision tree classifier model. Researchers have seen encouraging outcomes with this strategy. The AGD now gives more accurate findings than human auditing. An example of an AGD demonstration is provided in the appendix. The AGD system performance will be improved by posting it online for public usage and making use of user comments. The no projective dependency graph is included as a third element to improve the AGD's handling of Arabic odd permutation instances.

In 2021, researchers [17] worked on a rule-based Bengali Grammar Checker. A rapidly expanding area of artificial intelligence technology is computational linguistics, sometimes known as NLP. However, there has not been much progress made in this area of Bengali language. Spell checking, TTS, and optical character recognition (OCR) for Bengali have all been worked on. Researchers developed a method to verify the grammar of Bengali text since there has not been much effort done on Bengali grammar checking. Researchers drafted a few grammar rules that take into account the basic interdependencies of the various components of speech in order to examine the grammar of Bengali phrases. Researchers used pre-trained parts of speech (POS) taggers to break phrases into words before running our grammar software on them. This essay explained the use of POS tagging and grammatical rule checking to spot grammatical mistakes and misplaced words in Bengali phrases. By defining the sentence's parts of speech and then checking the sentence's structure against a set of rules, this approach identified structural grammatical faults in Bengali sentences. An accurate Bengali Grammar Checker is currently being implemented via a number of initiatives. An online platform for rule-based Bengali grammatical verification was the goal of this project. Even so, there are a number of tools available to deal with the processing aspect of NLP for Bengali grammar. No single platform verifies every rule to identify mistakes in Bengali phrases' grammar. This study tackled this problem and developed a platform that uses rules to verify Bengali sentence construction. The contribution of this study consists of categorizing words into seven parts of speech tags and entity rules for verifying nouns, adjectives, and verb phrases produced in accordance with Bengali grammar rules and programmed into the system to discover incorrect terms. In the future, a better POS tagger might be added to the system, greatly increasing its efficiency. The statistical approach might also be used to develop a recommendation system.

In 2021, this study [18] covered the continuing development of a deep neural network-based grammatical mistake detector for the Tamil language. This suggested grammar checker detected crucial subject-predicate agreement issues. In this instance, our focus is particularly on the agreement error between verbal predicates and nominal subjects. Additionally, researchers produced the first Tamil grammatical mistake annotated corpus ever. In order to capture syntactic information, the authors also used multilingual pre-trained language models. The researchers discovered that IndicBERT performed better on our tasks. With the use of our grammar-error annotated data, researchers improved the IndicBERT pre-trained model before implementing this grammar checker as a multi-class classification on top of it. The F1 score produced by this baseline model is 84.0. Researchers are now working to make this suggested approach better by using a dependency parser. Using a cutting-edge methodology, researchers have created a baseline purpose for Tamil grammatical mistake detection. The program described here looks for grammatical mistakes in sentences where the denominative subject and verbal predicate agree on the person number and gender. To capture the Tamil structures, researchers employed a multilingual pre-trained model. It was then adjusted using the grammatical error annotated data researchers produced. The authors discovered that compared to other pre-trained models, the IndicBERT model provides greater accuracy. For an unknown test set, our baseline model displays an F1 Score of 84.0%. To train the grammar checker, the researcher will produce additional annotated data using ThamizhiMorph, a morphological analyzer. To capture syntactic information like subject and predicate, the present model depends on a pre-trained model. A syntactic parser may be used to achieve this, and the syntactically parsed data may improve the score. As a result, as a further step, the researcher will test the suggested system by including syntactic information such as subject and predicate information into our datasets using a Tamil dependency parser named ThamizhiUDp.

In 2022, this paper [19] outlined the architecture and layout of the Nepali Grammar Checker, which is currently under study and improvement. The process is modular that has a grammar checker consisting of self-governing units. The units function as a pipeline to the in general system. Its grammar checker aims to check grammar errors, such as verbal and nominal agreement, parts of speech inflections, clause and phrase structure, and the different types of sentence structures for Nepali. The modules suggested in the system and architectural plan of Nepali Grammar Checker work on the investigate and the growth of the initial two modules, namely the tokenizer and the morphological analyzer, were completed. A prototype of both modules is also available and is being tested with additional features for difficulty handling. As previously talk about, the planned architecture and system concept for this system Nepali Grammar Checker is subject to modifications as the results of their research are made public in the future.

In 2023, this paper's researchers [20] introduced Vyakranly as a Hindi Translation and Grammatical Error Detection Toolkit specifically designed for the Indian language Hindi. The primary aims of Vyakranly include the identification and rectification of spelling errors in Hindi text, the detection and repair of grammar errors in Hindi sentences, as well as the translation of English to Hindi and Hindi to English texts. Notable aspects of our research include the identification of Hindi spelling, as well as the detection of corrections and grammatical errors. Finding historical literature in Hindi is difficult because it lacks digital information and has a more intricate morphology than English.

In 2024, researchers [21] gave a solution under consideration that integrates an improved orthography corrector algorithm with a DeepSpeech2 model architecture that utilizes bilingual encoder symbol derived from transformers and gated recurrent units. The method used in this study enhanced existing decoding algorithms, such as greedy or prefix beam search. It utilized post-processing methods particularly tailored for alterations in the Gujarati language. To train the model, we collected high-quality Gujarati speech data from several speakers, both male and female, using crowd-sourcing. This ensured that we applied the most optimum parameter values. Overall, there was an important 17.20% decrease in word error rate (WER). Furthermore, this research examined a range of analytical methodologies aimed at detecting mistakes that arise from diacritics, consonants, independent components, homophones, and half-conjugates. The enhancement of the Automated speech recognition (ASR) system's general efficacy was attained by the acquisition of a more profound understanding of the Gujarati language and the subsequent use of these strategies.

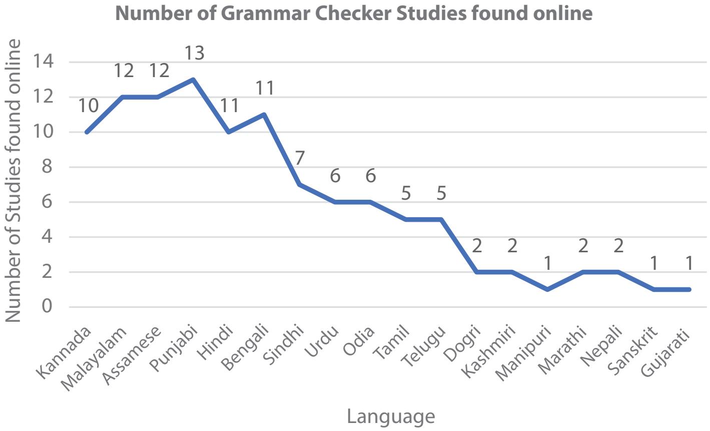


Figure 4.1 Indic language grammar checker research studies found online as of March 2024.

The graphical representation of the number of grammar checker studies found online up to March 2024 is shown in Figure 4.1. More than 109 research papers are available, all on purely Indian language grammar checkers. The highest number of research was found in South Indian languages compared to other Indian languages.

4.4 Spellchecker Related Work

In 2002, researchers [22] presented the Assamese Spellchecker are being considered. At first, the checker resorted to a dictionary lookup. If the dictionary did not have an entry for the term, a suggestion generator provided some replacements. After generating ideas using three different methods and ranking them, the final results were shown to the user. Spellchecker results were shown to be adequate in studies of manuscripts including over 5000 words, including juktaksharas. The Resource Centre for Indian Language Technology Solutions at the Indian Institute of Technology Guwahati is also looking into ways to include its Assamese to English online dictionary into the spellchecker software. Users may choose any of the two languages, type a word, and get the translation in the other. The spellchecker is being modified to work with the dictionary; currently, it exists as a Computer-generated imagery (CGI) software on the web server and its integration with the dictionary is under progress. Databases are made up of dictionaries, which include information on the term, including its definition, grammatical classification, pronunciation, pronunciation, antonyms, and synonyms. The performance of spell checking might be improved with the addition of an additional data set including the Soundex code of all Assamese words. Assamese bigram sequences have been gathered, and the list may be updated in the future. The code to identify spelling errors through a bigram search has also been built.

In 2012, the researchers [23] described the total design and implementation of a Kashmir Spellchecker and presented them this research paper. There were no things undertaken for the Kashmiri language, despite the fact that all main word processors present spell checking for English and in many European languages, with many Indian languages like Hindi, Urdu, Sanskrit, Tamil, and Kannada. The spellchecker for Kashmir that was developed is a standalone application that is not a component of any text editor. Researchers only corrected non-real word errors in this system. About 80% of errors are caught by the framework, and it offers 85% of the right suggestions. Future research will focus on the identification and correction of real word errors. This system can be used for other languages as well, but researchers should incorporate it as an add-on for Open Office and must first create a lexicon for those other language families.

In 2013 [31], this paper's researchers [24] discussed language authorities or consortiums that standardize word spellings, which are then made accessible in dictionaries or lexicons. For example, the word "produkt" is not included in the English dictionary. Therefore, in Urdu, "produkt" is a non-word but "produkt" is a properly spelt word. In today's digital world, text is often represented electronically. English is one of several languages with a variety of applications and additional tools. On the other hand, the creation of applications for less resource-intensive languages like Urdu is still in its infancy. When people compose material digitally on computers, spelling is crucial. The fact that terabytes of text are contributed in the form of corpora or other types of material that need to be spell checked and that doing so manually is very difficult is not taken into account. Numerous spell-checking methods were investigated and assessed in this paper. Spell checking may be done using each one of them independently or in combination. Reverse edit distance approach, a version of the edit distance technique, was used to recommend appropriate words for non-words in spellcheckers for many languages. For an Urdu word of length "n," candidates are selected by doing 86n+41 comparisons. Although the Urdu language has a rich literary history in South Asia, it lacks resources for computer-based projects. The spelling mistake detection and repair functionality in various electronic applications is the main emphasis of this study. The focus of this effort is on assembling a variety of spell-checking and error-correction methods that may be used to fix Urdu spelling mistakes. The difficulty of the reverse edit distance method is calculated to be 86n + 41. The method must be created for Urdu even if it has already been done for languages like English.

In 2015, the researchers [25] described Hindi Spellchecker. The three basic functions of this technology are error detection, error repair via providing recommendations, and error substitution. Approximately 83.2% of spelling mistakes are caught by the algorithm, which also offers 77.9% of the right word recommendations. The HINSPELL-Hindi spell checking system, which is not a feature of any word processor, is presented in this article. Only non-word mistakes are dealt with by this approach. Future studies will focus on actual grammatical faults. The system reports a detection rate of 83.2% and a correction rate of 77.9%. There is room for development in the deployment of SMT with lower reaction time as the correctness of the system grows but the system's answer time also increases after using SMTTechnique. HINSPELL may be adjusted for usage with various languages by changing the dictionary and keyboard.

In 2015, the researchers [9] described Tamil as a difficult inflectional language, which makes developing a morphological analyzer for it difficult for computer linguists. Tamil morphological generators and analyzers are being developed. Making a comprehensive and effective operating system is quite different from making a morphological generator or analyzer for demonstration purposes. Making a system that provides 60 to 65% coverage is not too difficult. It is challenging to raise efficiency over 65% and to a realistic level of 95% to 97%. Instead of being broad and detailed, the majority of the current descriptions concentrate on a vertical feature that is limited. The coverage that is attained by including the descriptions of the previous work is often never more than 50%. The assessment technique reveals that the effectiveness of the morphological analyzer created by the current team for Tamil is really impressive. The creation of a morphological analyzer, which includes the analysis of verbal complex, has several natural language applications, including lemmatization, text production, machine translation, and parsing. Additionally, it aids with document retrieval and word processing applications like spell checking and text input as well as voice applications like TTS synthesis and speech recognition. Assigning POS tags to the verbal complex is made easier by locating the categorical details of the verbal forms.

In 2016, the researchers [26] described a comprehensive plan and growth of a spellchecker for Kashmir and presented this as the research paper's main goal. There have not been anything undertaken for the Kashmiri language, despite the fact that all main word processors offer spell checking for English and in many European languages, along with many Indian languages like Hindi, Urdu, Sanskrit, Tamil, and Kannada. The spellchecker for Kashmir that was developed is a small program that is not a component of any word processor. Researchers have only corrected non-real word errors in this system. About 80% of errors are caught by the system, and it offers 85% of the right recommendations. Future research will focus on the identification and alteration of real word fault. This system can be used for other languages as well, but researchers will have enforce it as an add-on for Open Office and will need to start creating a lexicon to use it.

In 2016, the researchers for this study [27] claimed that a spellchecker is a tool to evaluate words in a text, check their spelling, and, if a spelling error is found, provide recommendations for replacement terms. Numerous studies on spellcheckers for Indian and European languages are available. However, there are not many for Tamil, maybe due to the language's inherent difficulty and great inflectional complexity. This research suggests an efficient method for identifying and fixing errors in Tamil spell checking. A new hybrid method is put into practice by combining n-gram and stemming techniques with a tree-based algorithm. The results of our testing demonstrate that our system can accurately identify spelling errors and provide the majority of workable options for repairing misspelled words with at least 91% accuracy. The performance of mistake detection and correction modules for the Tamil language was compared in this study. In this regard, several test word sets were used to evaluate and apply the tree-based method, n-gram, minimal edit distance, stemming, and lemmatization techniques. The results show that the tree-based algorithm is a much better method for error identification than the other two approaches and that the n-gram methodology, which uses stemmed words, provides the greatest proposal for fixing misspelled words. The results of our testing show that our system can accurately identify spelling errors and provide the best advice for correcting misspelled words with at least 91% accuracy. The n-gram technique is unable to suggest a repair for misspelled words since the accuracy decrease is associated with terms that are not included in the created corpus.

In 2016, this paper's researchers [28] described a program called "Spell Checker" which deals with spelling mistakes and spelling variations (SV). All misspelled words are noted and given the opportunity to be corrected. The content is examined for spelling mistakes and suggestions for repair are given when using this software as an edit. Telugu is an agglutinating language with a highly intricate morphology and a voluminous system of morphophonemic. Telugu has both internal and exterior sandhi, which is something that is noticeable. Telugu has a lot of vocalic and consonantal sandhi, both of which have been extensively examined. It is a very difficult process to determine the precise sandhi kind and separate it suitably. An assortment of linguistic modifications at word borders are known as external sandhi. These modifications resemble phonological processes like deletion, insertion, and substation. In Telugu, external sandhi is often mirrored orthographically. In these situations, external sandhi results in the development of forms that are morphologically impossible to analyze, creating a challenge for all types of NLP applications. In this essay, researchers go into great depth on the Telugu external sandhi procedures and the Spell Checker computer software.

In 2018, the suggested study [29, 37] was the first of its type and an innovative endeavor that concentrated on utilizing deep learning to construct a spellchecker for Malayalam. Two procedures made up the spellchecker: mistake detection and repair. An LSTM-based neural network is used in the error detection part and is trained to recognize misspelled words and the exact location where the mistake happened. The F1 score is used to calculate the error detection accuracy. By choosing the most likely term from the list of candidate words, errors may be corrected. One of the difficulties in doing this task is the lack of clear data. This paper used a simple neural network model. The use of sophisticated neural network topologies for experimentation was hampered by limited computing resources. Despite these restrictions, the suggested approach outperformed the conventional unicode splitting techniques. The outcomes may be significantly enhanced with access to massive clean corpora. Advanced neural network models [38, 39] still need improvement, and it is time to compare their performance to that of the current spell checking techniques. This investigation demonstrates that the word splitting used in the suggested strategy outperformed unicode splitting. Deep learning [40] was used to construct a spellchecker that is capable of both mistake detection and repair. The character level spellchecker for Malayalam is hopeful and merits more investigation.

In 2019, the researchers of this paper [30] prepared the primary objective of this paper which is to provide the best possible alternative spellings for Bengali terms and to identify instances where they have been misused. Any technique that utilizes the Bengali language may take advantage of the platform designed to verify spelling since it is a universal component that can be included into any system. Because of its intricate nature and grammatical constraints, developing an effective spellchecker for Bengali is usually a significant undertaking. The rule is frequently broken in Bengali, which is termed exceptional. In light of this, researchers offer an algorithm that is a clever hybrid of many well-known methods, such as edit distance and Soundex matching. Researchers can predict which incorrect Bengali words are most likely to be suggested after combining existed algorithms and running them through the decoding Bangla familiar to the algorithms.

In 2020, according to the paper's reviewer [31], the functionality of word processors, search engines, and social media platforms is heavily reliant on their grammar and spell checks. Spellcheckers are language-analysis programs that scan a text for spelling errors. If there are any accidental typos in the content, the user is alerted. Researchers are in need of a thorough investigation of the field of spell-checking, including aspects such as power, limitations, handled errors, performance, and evaluation standards. Spellcheckers for some languages are available in literature, and although they all have distinct designs, they all have similar features. This study applies the principles of a methodical literature re-examination of the domain of spell-checking. The methods of the methodical literature re-examination are used on 130 selected articles that were published in prestigious journals, conferences, and workshops related to the subject of spell checking more languages, and a wider variety of creations. During these stages, research questions are developed, research papers are chosen, inclusion and exclusion criteria are used, and pertinent data is extracted from the selected research articles. The major sub-areas of the literature on spell checking are categorized by language. Afterwards, each sub-area is covered in depth based on the method used. To get the outcome of this study, a number of articles are evaluated according to predetermined standards. This article explains how spellcheckers may be created using techniques from different domains, such as morphology, POS, chunking, stemming, and hash tables. It also describes the main problems that researchers ran across and the direction that spell-checking research will go in the future. A basic linguistic tool of NLP, spellcheckers are used in email clients, social media, search engines, information extraction, proofreading, and information retrieval, among other uses. In order to clearly distinguish between them, the spellcheckers of any language are evaluated based on the scripts, approach used, available tools, corpus size, performance, and error handled. The comparative research indicated above also helps choose the spellchecker that fits the language category in question the best in terms of accuracy and ease of use. Of all the spellcheckers available, the Assamese language's categorization and confusion-set-based spellchecker is the finest. Nonetheless, there are two notable spellcheckers available for distinct scripts in the Punjabi language group, and they are both rather effective. Additionally, the n-gram-based spell checking performs very well in the Tamil language category, while the Medical (MED)-based spellchecker carry out best in the Urdu language area. Existing spellcheckers for Telugu are based on depth algorithms and statistical techniques. The spellchecker based on deep learning performs better in this language area. The n-gram and MED based spellchecker for Malayalam produce the same results as other spellcheckers available today. Moreover, there is a single spellchecker for other languages like Sindhi, Odia, and Kashmiri, and it works really well. This article also describes the use of rule-based, statistical, and deep learning technologies in the spell-checking industry. The rule-based techniques do check for spelling by using heuristics based on chunking, stemming, morphology, and POS. The statistical methods use word counts, word frequencies, and word characteristics for spell checking. While statistical and rule-based techniques are quite effective in identifying non-word errors, their applicability to real-world errors is restricted. Spellcheckers based on deep learning are necessary to manage such errors. These techniques work rather well in preventing real-world errors. Although research on magical deep learning methods is still in its early stages, the encouraging results thus far suggest that similar techniques might be useful in the future for other languages to enhance spell-checking performance. This article also describes the challenges faced by the researchers, such as the complexity of time and space, the inability to distinguish words in certain scripts that are context-sensitive, and having to understand context-sensitive spelling.

In 2020, the researchers [32] described that the lives of a sizable portion of computer users are made possible by the unseen but essential role played by spellcheckers. The majority of text-manipulating software, including office packages with anything from basic note-taking tools to fully functional word processors and search engines, are included with it. Text is tokenized by spellcheckers to check for grammatical and spelling mistakes. Although they function for the majority of European languages, existing spellcheckers do not take into consideration the linguistic characteristics of Indian languages like Tamil. There is an additional complication to how Tamil text should be treated for tools like spell checking due to problems with Tamil's unicode encoding representation. Three alternative spellcheckers—bloom-filter, symspell, and LSTM-based—are implemented and tested for Tamil. For lookup recommendations and validation, Symspell is quite quick. Although not precise enough for daily usage, LSTM implementation is an intriguing area of research that has not yet been fully explored. Despite the fact that it is difficult to develop a fully automated spellchecker, it is worthwhile to try to make one since spelling and grammar checking is essentially an Artificial general intelligence (AGI) issue. The current spellcheckers operate in various situations using a variety of different criteria. Grammarly and other such programs are excellent for English. Due to its current infrastructure, it has daily access to a vast quantity of data that it can use to learn from and enhance its algorithms. Researchers look forward to creating such a platform for Tamil with the help of the collective effort from the Tamil speaking community, even if our endeavor is just an initial step in that regard.

In 2021, the researchers of this paper [33] prepared and introduced a Gujarati spellchecker tool 'Jodani,' which is a mechanism on root words and employs string similarity metrics to recognize incorrectly spelled words before attempting to auto-correct the word or suggest syntactically appropriate terms. Jodani determines the proper spelling of a Gujarati word and returns a list of recommendations for the wrong word. The spelling accuracy and list of suggestions are created using the Levenshtein edit distance, and procedures for handling inflected words and sorting the suggestions are devised. The accuracy of 91.56% demonstrates the effectiveness of Jodani's method, which outperforms the existing Gujarati spellchecker. Further improvements can be made by overcoming the assumption that the initial character of the entered word is valid. Machine learning approaches, such as the neural network, can also be used to enhance the order of suggestions.

In 2022, this paper's researchers [34] discussed that a spellchecker is a piece of software that indicates words in a text that could be misspelled. Regardless of the language, a spellchecker is a necessary function of any word processor. The spellchecker looks for misspelled words in the written text and offers the finest correction options. In this study, a hybrid design and implementation method was used for the first time to create a Dogri spellchecker. The main parts of this system are error detection, suggestion-based error correction, and manual or automated error replacement following the suggested technique. The system discovered 80.79% of misspelled words and accurately proposed 74.72% of incorrect words.

In 2023, this paper's researchers [35] found an analysis of the construction of a model aimed at identifying incorrect Assamese words inside digital information. When deciding the spelling of a word, the proposed model considered the contextual information inside the sentence. The comment underscores the fact that many Assamese words possess several interpretations and may not align with the intended context of a given remark while being written accurately as per the dictionary. The chosen methodology utilized two distinct ML methodologies, namely LSTM and bidirectional long short-term memory (BiLSTM). The findings indicated that the suggested move towards significantly improving the identification of spelling errors in the Assamese language, surpassing the presentation of other studies. The BiLSTM model achieved the highest accuracy of 89.52% in the suggested strategy.

In 2024, This research [36] aimed to remedy the lack of a reference point for the job by creating a comprehensive parallel corpus consisting of 7.7

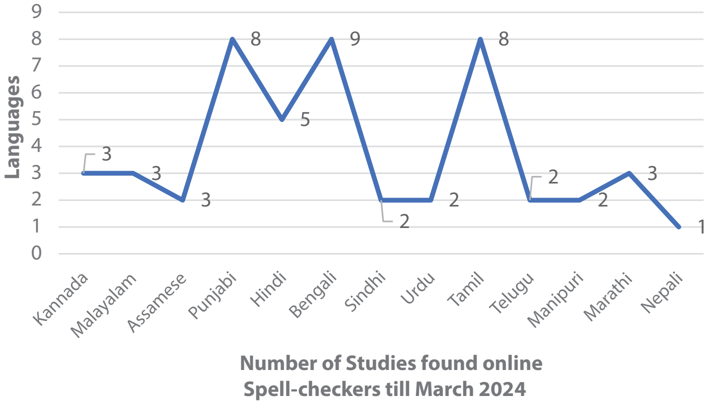


Figure 4.2 Indic language Spellchecker research studies found online as of March 2024.

million source-target pairs. Additionally, we investigated the unexplored capabilities of transformers. In addition to the corpus, we presented a novel approach called Panini, which utilizes transfer learning to achieve well-organized, monolingual, transformer-based correction of Bangla grammatical errors. This method has gained recognition as the leading approach for this task, outperforming both BanglaT5 and T5-Small by 18.81% and 23.8% in terms of accuracy scores, and 11.5 and 15.6 in terms of SacreBLEU scores, respectively. The technique's empirical results provide evidence supporting its superiority over other techniques in capturing complex language rules and patterns. Furthermore, we compared the effectiveness of our suggested approach to the Bangla paraphrase task. This comparison demonstrates its higher performance, surpassing the previous cutting-edge technique in this work. The BanglaGEC corpus and Panini, as well as the baselines for BGEC and the Bangla paraphrase task, are openly available resources.

A graphical depiction of the number of spellchecker studies available online as of March 2024 can be seen in Figure 4.2. There are more than 50 research publications available that focus just on spellchecking in the Indian language. Bengali and Punjabi languages have the most study findings when compared to other Indian languages.

4.5 Conclusion

Overall, this study describes the current scenario of an NLP-based spell and grammar checker system for Indian languages. A total of six types of spell and grammar checker development techniques are explained very briefly. Afterwards, researchers focus on in-depth spell and grammar checker studies with a critical explanation of each paper. Through that critical discussion, one can find the research gap in the current system. If one can work on Indian languages based on NLP research, this system may help them elaborate on their research in the right way. The main issue of Indic languages NLP research is the limited resources of each language, mainly datasets. One can find his or her own way to the next NLP research path through this study in the context of Indian languages only.

References

1. Kabir, H., Nayyer, S., Zaman, J., Hussain, S., Two pass parsing implementation for an Urdu grammar checker, in: *Proceedings of IEEE international multi topic conference*, pp. 1–8, 2002.
2. Kabir, H., Two-pass parsing implementation for an Urdu Grammar Checker. *International Multi Topic Conference, 2002. Abstracts. INMIC 2002*, pp. 51–51, 2002, doi: 10.1109/INMIC.2002.1310158.
3. Bal, B.K., Shrestha, P., Pustakalaya, M.P., Dhoka, N.P., Architectural and system design of the Nepali grammar checker, in: *PAN localization working paper*, 2007.
4. Gill, M. and Gurpreet, L,., A Grammar Checking System for Punjabi. *IEEE – International Conference*, 1, 149–152, 2008.
5. Gill, M.S., Lehal, G.S., Joshi, S.S., Part-of-speech tagging for grammar checking of Punjabi. *Linguist. J.*, 8, 6–22, 2009.
6. Bopche, L. and Dhopavakar, G., Rule Based Grammar Checking System for Hindi. *J. Inf. Syst. Commun.*, 3, 1, 45–47, 2012, ISSN: 0976-8742 & E-issn: 0976-8750.
7. Kaur, J. and Kamal, G,., Hybrid Approach for Spell Checker and Grammar Checker for Punjabi. *Int. J. Comput. Sci. Software Eng.*, 4, 6, 62–67, June 2014.
8. Sarma, H., Das, D., Kashyap, K., Grammatical Error Detection Model for Assamese Sentences. *Int. J. Adv. Res. Comput. Commun. Eng.*, 4, 11, 529–531, 2015.
9. Sankaravelayuthan, R., Spell and grammar checker for Tamil', 1, Wiley, 2015, 10.13140/RG.2.1.3700.6803.
10. Sharma, and Lehal, G.S., Improving existing Punjabi Grammar Checker. *2016 International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT)*, pp. 445–449, 2016, doi: 10.1109/ICCTICT.2016.7514622.
11. Sharma, S.K. and Lehal, G.S., Improving existing Punjabi Grammar Checker. *2016 International Conference on Computational Techniques in Information*

and Communication Technologies (ICCTICT), New Delhi, India, pp. 445–449, 2016, doi: 10.1109/ICCTICT.2016.7514622.

12. Schmaltz, A., Alexander, Y.K., Rush Stuart, M., Shieber, M., Adapting Sequence Models for Sentence Correction. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, pp. 2807–2813, Association for Computational Linguistics, Harvard University, September 7–11, 2017.
13. Patel, N. and Patel, D., Research review of Rule Based Gujarati Grammar Implementation with the Concepts of Natural Language Processing (NLP), 1, Wiley, 2018, 10.6084/m9.jetir.JETIRA006276.
14. Mittal, M., Sharma, S.K., Sethi, A., Detection and Correction of Grammatical Errors in Hindi Language, Using Hybrid Approach. *Int. J. Comput. Sci. Eng.*, 7, 5, 421–426, May 2019, E-issn: 2347-2693.
15. Hulipalled, B. C, V. R. and Simha, J.B., Kannada Grammar Checker Using LSTM Neural Network. *2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE)*, Bengaluru, India, pp. 332–337, 2020, doi: 10.1109/ICSTCEE49637.2020.9277479.
16. Alothman, A. and Alsalman, A.M., An Arabic Grammar Auditor Based on Dependency Grammar. *Adv. Hum. Comput. Interact.*, 2020, 10 pages, 2020, Article ID 8856843 https://doi.org/10.1155/2020/8856843.
17. Fahim Faisal, A.N.M., Rahman, M.A., Farah, T., A Rule-Based Bengali Grammar Checker. *2021 Fifth World Conference on Smart Trends in Systems Security and Sustainability (WorldS4)*, London, United Kingdom, pp. 113–117, 2021, doi: 10.1109/WorldS451998.2021.9514031.
18. Murugesapillai, D., Ravinthirarasa, A., Dias, G., Sarveswaran, K., Neural-based Tamil Grammar Error Detection, in: *Proceedings of the First Workshop on Parsing and its Applications for Indian Languages*, NIT Silchar, India, NLP Association of India (NLPAI), pp. 27–32, 2021.
19. Bal, B.K., Shrestha, P., Madan, P., Pustakalaya, Patandhoka, Nepal, Architectural and System Design of the Nepali Grammar Checker, 1, Wiley, 2022.
20. S., R., S., V., T., S., K., R., Gadhikar, L., Vyakranly : Hindi Grammar & Spelling Errors Detection and Correction System. *2023 5th Biennial International Conference on Nascent Technologies in Engineering (ICNTE)*, Navi Mumbai, India, pp. 1–6, 2023, doi: 10.1109/ICNTE56631.2023.10146610.
21. Bhagat, B. and Dua, M., Improved spell corrector algorithm and deepspeech2 model for enhancing end-to-end Gujarati language ASR performance, e-Prime. *Adv. Electr. Eng. Electron. Energy*, 7, 100441, 2024, ISSN 2772-6711, https://doi.org/10.1016/j.prime.2024.100441.
22. Das, M., Borgohain, S., Gogoi, J., Nair, S.B., Design and implementation of a spell checker for Assamese, in: *Language Engineering Conference, 2002. Proceedings*, IEEE, pp. 156–162, 2002.
23. Lawaye, A.A. and Purkayastha, B.S., Kashmiri spell checker and suggestion system. *Communications*, 21, 2, 123, 2012.

24. Iqbal, S., Anwar, W., Bajwa, U., II, Rehman, Z., Urdu spell checking: Reverse edit distance approach, in: *Proceedings of the 4th workshop on south and southeast asian natural language processing*, pp. 58–65, 2013.
25. Mishra, A. K., Tyagi, A. K., Dananjayan, S., Rajavat, A., Rawat, H., Rawat, A., Revolutionizing Government Operations: The Impact of Artificial Intelligence in Public Administration, in: *Conversational Artificial Intelligence*, pp. 607–634, 2024.
26. Lawaye, A. and Purkayastha, Design and Implementation of Spell Checker for Kashmiri. *Int. J. Sci. Res.*, 5, 199–200, 2016.
27. Sakuntharaj, R. and Mahesan, S., A novel hybrid approach to detect and correct spelling in Tamil text. *2016 IEEE International Conference on Information and Automation for Sustainability (ICIAfS)*, Galle, Sri Lanka, pp. 1–6, 2016, doi: 10.1109/ICIAFS.2016.7946522.
28. Uma Maheshwar Rao, G. and Amba, P., *Kulkarni, Christopher Mala & Parameshwari K., 2012*, Telugu Spell-Checker. Vaagartha, Editors- Shivarama Padikkal Tariq Khan, p. 57, 2022.
29. Sooraj, S., Manjusha, K., Kumar, M., Soman, K.P., Deep learning based spell checker for Malayalam language. *J. Intell. Fuzzy Syst.*, 34, 1427–1434, 2018, 10.3233/JIFS-169438.
30. Saha, S., Tabassum, F., Saha, K., Akter, M., Bangla spell checker and suggestion generator PhD diss, United International University, USA, 2019.
31. Singh, S. and Singh, S., Systematic review of spell-checkers for highly inflectional languages. *Artif. Intell. Rev.*, 53, 4051–4092, 2020, https://doi.org/10.1007/s10462-019-09787-4.
32. Murugan, S., Bakthavatchalam, T.A., Sankarasubbu, M., *SymSpell and LSTM based Spell-Checkers for Tamil*, Wiley - USA, 2020.
33. Patel, H., Patel, B., Lad, K., Jodani: A spell checking and suggesting tool for Gujarati language. *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pp. 94–99, 2021, doi: 10.1109/Confluence51648.2021.9377072.
34. Jamwal, S.S. and Gupta, P., A Novel Hybrid Approach for the Designing and Implementation of Dogri Spell Checker, in: *Data, Engineering and Applications. Lecture Notes in Electrical Engineering*, vol. 907, S.L. Peng, J. Agrawal, R.K. Shukla, D.N. Le (Eds.), Springer, Singapore, 2022, https://doi.org/10.1007/978-981-19-4687-5_53.
35. Phukan, R., Neog, M., Baruah, N., A Deep Learning Based Approach For Spelling Error Detection In The Assamese Language. *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Delhi, India, 2023, doi: 10.1109/ICCCNT56998.2023.10306972.
36. Hossain, N., Bijoy, M.H., Islam, S. *et al.*, Panini: a transformer-based grammatical error correction method for Bangla. *Neural Comput. Applic.*, 36, 3463–3477, 2024, https://doi.org/10.1007/s00521-023-09211-7.

37. Rathi, M. and Rajavat, A., Investigations and Design of Privacy-Preserving Data Mining Technique for Secure Data Publishing. *Int. J. Intell. Syst. Appl. Eng.*, *11*, 9s, 351–367, 2023.
38. Dubey, P. and Rajavat, A., Effective K-means clustering algorithm for efficient data mining, in: *2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN)*, IEEE, pp. 1–6, 2023, May.
39. Shrivastava, M., Patil, R., Bhardwaj, V., Rawat, R., Telang, S., Rawat, A., Quantum Computing and Security Aspects of Attention-Based Visual Question Answering with Long Short-Term Memory. *Quantum Comput. Cybersecur.*, 1, 395–412, 2023.
40. Rawat, R., Chakrawarti, R.K., Raj, A.S.A., Mani, G., Chidambarathanu, K., Bhardwaj, R., Association rule learning for threat analysis using traffic analysis and packet filtering approach. *Int. J. Inf. Technol.*, *15*, 6, 3245–3255, 2023.

5

Identification of Gujarati Ghazal Chanda with Cross-Platform Application

Brijeshkumar Y. Panchal

Computer Engineering Department, Sardar Vallabhbhai Patel Institute of Technology (SVIT)-Vasad, Gujarat Technological University (GTU), Anand, Gujarat, India

Abstract

Ghazal is a well-known poetry form of Indian literature. The Ghazal genre is centuries old, dating back to seventh-century Arabic poetry. In the 12th century, the Ghazal moved throughout South Asia. A Ghazal is a lyrical representation of both the anguish of loss or estrangement and the pleasure of love in the midst of that pain. As with many other languages, Guajarati Ghazal has numerous types of Gujarati Ghazal "Chanda". In this chapter, the researcher attempted to identify Gujarati Ghazal "Chanda" in the form of percentage using an android application and website platform. This researcher used the concept of Unicode, by inputting the string and converting it into characters. Then, their Unicode is generated and pushed into the stack. The output is compared with the fixed Unicode defined for the Gujarati character; if they matched, then the input in the new string will be in the form of 1's and 2's, which are then compared within the HashMap, and if they matched, it will give the output for the same simple Unicode. In the current scenario with few research gaps, the system has 72% accuracy for the identification of Gujarati Ghazal Chanda. The Gujarati language history, Ghazal history, and how to attempt a Ghazal are the contents of this chapter.

Keywords**:** Identification of chanda, Gujarati, Ghazal, chanda, cross-platform, identification chanda, poetry

Email: panchalbrijesh02@gmail.com

Rajesh Kumar Chakrawarti, Ranjana Sikarwar, Sanjaya Kumar Sarangi, Samson Arun Raj Albert Raj, Shweta Gupta, Krishnan Sakthidasan Sankaran and Romil Rawat (eds.) Natural Language Processing for Software Engineering, (71–94) © 2025 Scrivener Publishing LLC

Abbreviations

Ghazal	ગઝલ
Chand	છંદ
Matlaa	મતલા
Radif	રદફિ
Qaafiyaa	કવ્વાલી
Maqtaa	મક્તા
Beher	બહેર
Misra	મશિરાઆ
Śēra	શેર

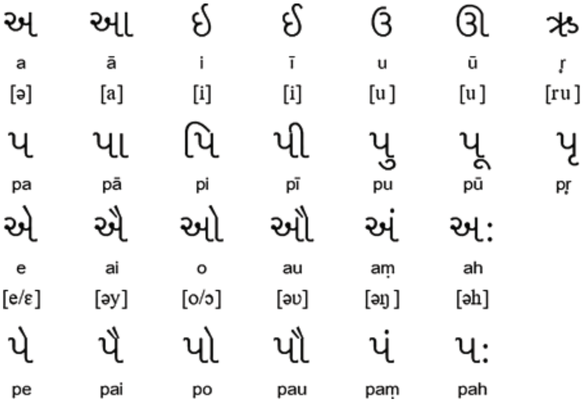
5.1 Introduction

5.1.1 The Gujarati Language

Gujarati is classified as an element of the southwestern part of the New Indo-Aryan language family, which is a subdivision of the Indo-Iranian stem of the Indo-European language relatives [1]. Gujarati is the official language of the Gujarat state and is widely spoken in various regions of South Asia, including Maharashtra, Rajasthan, Sind, lower Punjab, Madhya Pradesh, and Karnataka [2]. The Parsi, Hindu, Muslim, and Jain communities in the Persian Gulf, East and South Africa, Britain, North America, and Australia also speak Gujarati. Gujarati is the sixth most prevalent language in India in terms of resident speakers, with 55.5 million individuals speaking it, accounting for around 4.5% of the overall Indian population as of 2011. As of 2007, the number of native speakers ranked it as the 26th most commonly used tongue on a global scale. Subjects of study include history and literature [3]. Old Gujarati, Middle Gujarati, and Modern Gujarati are three distinct periods in the history of the Gujarati language. Old Gujarati refers to the period from the 12th to the 15th centuries; Middle Gujarati refers to the period from the 15th to the 18th centuries; and Modern Gujarati refers to the period from the 18th century onward [4].

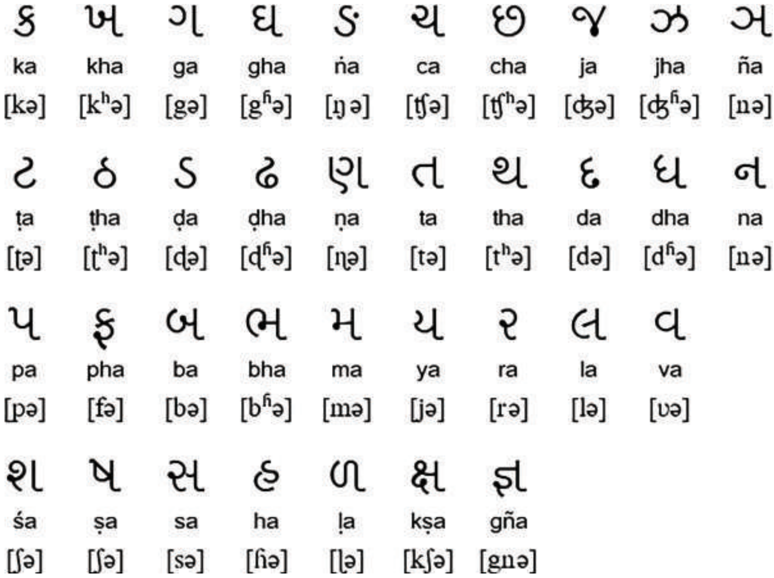
The origins of this may be traced back to a specific literary form in Old Western Rajasthani, even though there is evidence of Jain Prakrit discourse and studies by Middle Indian grammarians of Nagara Apabhramsa, a literary Apabhramsa of Gujarat. The Bharates-varabahubaliräsa, composed in 1185, is the oldest known Gujarati literary work from the 12th century [5]. There are collections of prose and poetry that were composed starting in the 13th century. These collections consist of the regular poem Vasantaviläsa and the 14th-century commentary called the Sadavaśyakabalabodhavrtti. Narasimha Mehta's (c. 1414–1488) devotional songs ushered in a new period in poetry and earned a prominent position in its literary history. *Mumbai Samăcar*, a Gujarati daily founded in 1822, is among the most ancient newspapers in Asia. The Bombay Parsis engages in recreation a pioneering role in the growth of Gujarati and Urdu theater [6] as shown in Figure 5.1, which illustrates the vowels of the Gujarati language [7].

Dialects of Gujarati, spoken along the Baroda-Ahmedabad Corridor, is considered as the standard or prestigious dialect [7]. The question is whether the list of Nagari Brhma's carries the same meaning. The position of recurrent platform ("RP") is subject to debate. Additional variations of the Gujarati language are Surati, Carotari, Käthiwārī, and Pātānī, which correspond to the southern, central, Saurashtra, and northern regions of Gujarat, respectively [8]. Pakistani Gujarati is likely a subdialect of the Păţăni language, and the practice of code flipping is declining as the younger population transitions to Urdu and regional languages [9]. Speakers of the Muslim faith, both in that location and in other places, clearly use a significant number of words from the Perso-Arabic language, which is their second-greatest source of vocabulary after Sanskrit. This is particularly evident in discussions related to religion and culture. Parsi Gujarat, a kind of language used by the Zoroastrians of the Indian subcontinent, is easily understood. Swahili loanwords have been incorporated into the East African Gujarat language. Kacchi (Kachchi) is a language that is in the semantic middle ground between Gujarati and Sindhi, and it is also affected by Marwārī [10] as shown in Figures 5.2 and 5.3, which illustrate the consonants and conjunct consonants of the Gujarati language, respectively [7].



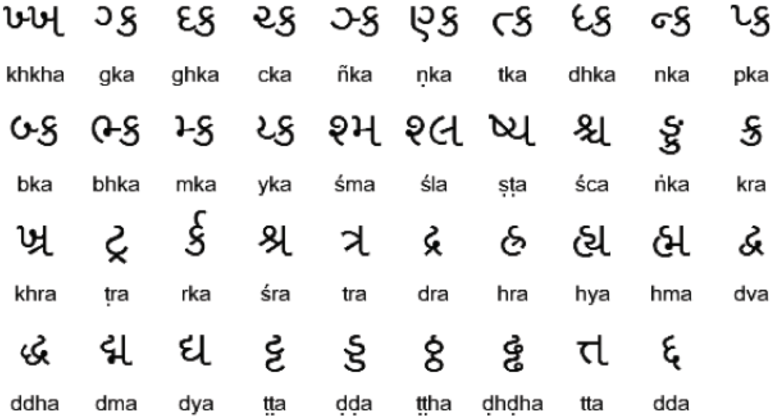
અ	આ	ઇ	ઈ	ઉ	ઊ	ઋ
a	ā	i	ī	u	ū	ṛ
[ə]	[a]	[i]	[i]	[u]	[u]	[ru]
પ	પા	પિ	પી	પુ	પૂ	પૃ
pa	pā	pi	pī	pu	pū	pṛ
એ	ઐ	ઓ	ઔ	અં	અ:	
e	ai	o	au	aṃ	ah	
[e/ɛ]	[əy]	[o/ɔ]	[əʋ]	[əŋ]	[əh]	
પે	પૈ	પો	પૌ	પં	પ:	
pe	pai	po	pau	paṃ	pah	

Figure 5.1 Vowels of the Gujarati language [7].



ક	ખ	ગ	ઘ	ઙ	ચ	છ	જ	ઝ	ઞ
ka	kha	ga	gha	ṅa	ca	cha	ja	jha	ña
[kə]	[kʰə]	[gə]	[gʱə]	[ŋə]	[ʧə]	[ʧʰə]	[ʤə]	[ʤʱə]	[ɲə]
ટ	ઠ	ડ	ઢ	ણ	ત	થ	દ	ધ	ન
ṭa	ṭha	ḍa	ḍha	ṇa	ta	tha	da	dha	na
[ʈə]	[ʈʰə]	[ɖə]	[ɖʱə]	[ɳə]	[tə]	[tʰə]	[də]	[dʱə]	[n̪ə]
પ	ફ	બ	ભ	મ	ય	ર	લ	વ	
pa	pha	ba	bha	ma	ya	ra	la	va	
[pə]	[fə]	[bə]	[bʱə]	[mə]	[jə]	[rə]	[lə]	[ʋə]	
શ	ષ	સ	હ	ળ	ક્ષ	જ્ઞ			
śa	ṣa	sa	ha	ḷa	kṣa	gña			
[ʃə]	[ʃə]	[sə]	[ɦə]	[ɭə]	[kʃə]	[gɲə]			

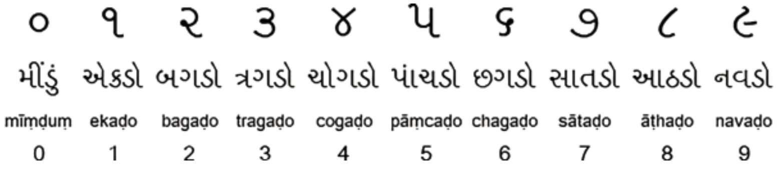
Figure 5.2 Consonants of the Gujarati language [7].



ખ્ખ	ગ્ક	ઘ્ક	ચ્ક	ઞ્ક	ણ્ક	ત્ક	ધ્ક	ન્ક	પ્ક
khkha	gka	ghka	cka	ñka	ṇka	tka	dhka	nka	pka
બ્ક	ભ્ક	મ્ક	ય્ક	શ્મ	શ્લ	ષ્ટ	શ્ચ	ઙ્ક	ક્ર
bka	bhka	mka	yka	śma	śla	ṣṭa	śca	ṅka	kra
ખ્ર	ટ્ર	ર્ક	શ્ર	ત્ર	દ્ર	હ્ર	હ્ય	હ્મ	દ્વ
khra	ṭra	rka	śra	tra	dra	hra	hya	hma	dva
દ્ધ	દ્મ	દ્ય	ટ્ટ	ડ્ડ	ટ્ઠ	ડ્ઢ	ત્ત	દ્દ	
ddha	dma	dya	ṭṭa	ḍḍa	ṭṭha	ḍhḍha	tta	dda	

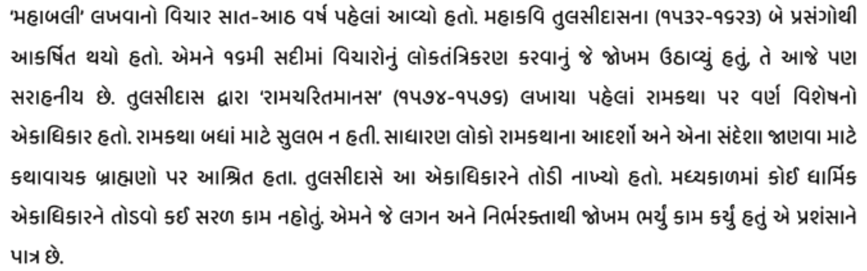
Figure 5.3 Conjunct consonants of the Gujarati language [7].

Gujart Grammar's phonetics are distinct due to the presence of muttered vowels that have evolved from the last /h/ sound, as well as two open vowels, /e/ and /5/. There is no difference in the length of the /i/ and /u/ vowels. Nouns, adjectives, and pronouns may be variable or invariable. There are three genders, counting the neuter, and two numbers. Direct and oblique forms undergo inflection, with the latter using post-positions



૦	૧	૨	૩	૪	૫	૬	૭	૮	૯
મીંડું	એકડો	બગડો	ત્રગડો	ચોગડો	પાંચડો	છગડો	સાતડો	આઠડો	નવડો
mīṃḍuṃ	ekaḍo	bagaḍo	tragaḍo	cogaḍo	pāṃcaḍo	chagaḍo	sātaḍo	āṭhaḍo	navaḍo
0	1	2	3	4	5	6	7	8	9

Figure 5.4 Numerals of the Gujarati language [7].



'મહાબલી' લખવાનો વિચાર સાત-આઠ વર્ષ પહેલાં આવ્યો હતો. મહાકવિ તુલસીદાસના (૧૫૩૨-૧૬૨૩) બે પ્રસંગોથી આકર્ષિત થયો હતો. એમને ૧૬મી સદીમાં વિચારોનું લોકતંત્રિકરણ કરવાનું જે જોખમ ઉઠાવ્યું હતું, તે આજે પણ સરાહનીય છે. તુલસીદાસ દ્વારા 'રામચરિતમાનસ' (૧૫૭૪-૧૫૭૬) લખાયા પહેલાં રામકથા પર વર્ણ વિશેષનો એકાધિકાર હતો. રામકથા બધાં માટે સુલભ ન હતી. સાધારણ લોકો રામકથાના આદર્શો અને એના સંદેશા જાણવા માટે કથાવાચક બ્રાહ્મણો પર આશ્રિત હતા. તુલસીદાસે આ એકાધિકારને તોડી નાખ્યો હતો. મધ્યકાળમાં કોઈ ધાર્મિક એકાધિકારને તોડવો કઈ સરળ કામ નહોતું. એમને જે લગન અને નિર્ભરક્તાથી જોખમ ભર્યું કામ કર્યું હતું એ પ્રશંસાને પાત્ર છે.

Figure 5.5 Sample text in Gujarati.

and clitics [11]. Verbal forms exhibit distinctions in terms of time, modality, and aspect. Verbal nouns and adjectives with auxiliaries create a wide range of obligation and desiderative forms. The vocabulary is abundant in passive, causative, and double causative verbs (Cardona, 1965). Vector or compound verbs, which are often found in New Indo-Aryan languages, are used in limited situations with precise meanings [12] as shown in Figures 5.4 and 5.5, which present the numerals of the Gujarati language and sample text in Gujarati, respectively [7].

5.2 Ghazal

Ghazal is considered the core part of the people of Gujrat. Upon analysis, researchers of the current market noticed that no applications have been developed especially to find Chanda of Gujarati Ghazal. While discussing with various literature experts, this researcher came to know that this is a good initiative to start because new learners who will choose this field. This researcher had a chance to contribute to our native language [13].

Here, this researcher used the concept of the Unicode for the identification of Ghazal Chanda. What is Unicode? Unicode is a standard for computers created to reliably and individually encode characters found in all written languages worldwide. The Unicode standard uses hexadecimal to represent characters. For instance, the Latin letter A is defined by the number 0x0041. Thus, this researcher used this concept, by inputting the string and converting it into characters. Then, their Unicode is generated and pushed into the stack. The output is then compared with the fixed Unicode defined for the Gujarati character, and if they matched, then the input the new string will be in the form of 1's and 2's, which are then compared within the HashMap, and if they matched, it will give the output for the same Unicode. The second thing this researcher did for this project involves the accuracy calculator: inputting the string follows the same process of the Unicode method, and if it matches, it will give the accuracy percentage as output [14].

Furthermore, this researcher used a number of modules for performing the same operation to match the string with the HashMap table. This researcher also used the stack for inputting the data; if any Matra comes out after popping, it will give an output of 2 depending on its strength; otherwise, it will give an output of 1 for the single character as well as for the Matra with a lower strength, which helps us find the Chanda of the string.

Orthograph, a work from 1592 (Mistry, 1996), confirms that a script based on a version of Devanagari has been used to write Gujarat and Kacchi since the 16th century. When printing was introduced in the 1830s, a cursive style took the place of the traditional Sanskrit script that was previously employed in both prose and poetry. There are a total of 45 symbols used to indicate independent and conjunctive forms [14]. These symbols include 8 vowels, 34 consonants, anusvära, visarga, and a velar nasal grapheme. Gujarati, a language written from left to right, is notable for its lack of head strikes and its use of different phonemic alterations. Similar to other characters inherited from Brahmi, it is clear that a post-consonantal /a/ is implied in a consonant that does not include diacritics. The numbers originating from Devanagari were adopted, although the forms of the digits 3, 5, 6, and 9 were changed.

One can create a Ghazal in Gujarati by following these simple guidelines [15]:

1. Select a subject or topic for the Ghazal. This might be anything that inspires you, whether it is spirituality, nature, or love.
2. Put your Ghazal together with a sequence of topically as well as emotionally related couplets, or "sher." Every couplet

needs to be able to stand on its own and express a whole idea.

3. Continue the Ghazal using the same rhyming pattern. The most typical pattern in Gujarati ghazals is AA, BA, CA, and so on, with the rhyme scheme being the same in both lines of each couplet.
4. In every couplet, use the radif, a word or phrase that repeats at the end of the second line. The Ghazal is more cohesive and has a feeling of flow because of its refrain.
5. Through the couplets, convey your feelings, ideas, and visuals while eloquently and poetically expressing the core of the selected topic.
6. The Ghazal's meter and rhythm should be observed since they enhance its melody and impact.

These instructions will help you write a stunning and moving Gujarati Ghazal.

5.3 History and Grammar of Ghazal

The Ghazal form has been around for centuries, going all the way back to seventh century Arabic poetry. The popularity of Ghazal increased over South Asia in the eleventh century. A Ghazal is a poetic look of the grief of loss or alienation, as well as the joy of love in the midst of such misery. Now, focus on Chanda; what is Chanda? It is the study of poetic meters and verse from numerous Gujarati Ghazalkar; learners may understand the structure of Ghazal and its five rules, namely, "Matlaa", "Radif", "Qaafiyaa", "Maqtaa", and "Beher" [16]. The basic unit of ghazal is Śēra. A ghazal is made up of three or more stocks. The shares of a ghazal are connected by the same rhyme, the same type of kafia, and the same verse. Two rows of a share are called two Misra. The first syllable of the ghazal is called Matla, both of which have to maintain the Radif-Kafia scheme in Misra. The Radif-Kafia scheme established in Matla then has to be carried out throughout the ghazal. Matla can be more than one. Shares other than Matla do not have a Radif in the first Misra. In the second Misra, the plan of Radif-Kafia has to be maintained [17, 18] as given in Table 5.1. Understanding Matra with a poetic line is shown in Table 5.2.

In Ghazal, researchers can identify Ghazal Chanda in a word-to-word and a character-to-character basis. Note that if one character has 1 matra, it means it is લ; if there are 2 matras, it means it is ગા.

Table 5.1 Understanding Matra.

1 = લ
2 = ગા

Table 5.2 Understanding Matra with a poetic line.

એ	ક	ધા	રી	ફૂ	લ	ની	કે	વી	અ	દા	છે
ગા	લ	ગા	ગા	ગા	લ	ગા	ગા	ગા	લ	ગા	ગા
2	1	2	2	2	1	2	2	2	1	2	2

Take an example of Ghazal with a few Chandas:

(1) એકધારી ફૂલની કેવી અદા છે.-
ગાલગાગા / ગાલગાગા / ગાલગાગા-
2122/2122/2122 - ખફીફ છંદ.
(2) ઝરુખે ખીલી છે ઉદાસી –
લગાગા / લગાગા / લગાગા – 1
122 / 122 / 122 - મુતકારબિ છંદ
(3) સાથ ચાલી દુઃખી સૌને કરવા હતા-
ગાલગા / ગાલગા / ગાલગા / ગાલગા-
212 / 212 / 212 / 212 મુતદારકિ છંદ.

5.4 Literature Review

The study on part-of-speech tagging for Gujarati using conditional random fields was suggested by Chirag Patel and Karthik Gali in 2008. A machine learning approach for Gujarati part-of-speech tagging has been presented by the author. The machine learning portion is finished. A cross-platform (CRF) model is a useful tool to use. This is taken into consideration when selecting the attributes that CRF receives. Considering the linguistic component of Gujarati, Gujarati now has a lower position than other languages in terms of resource scarcity. There are only around 600 manually marked records. There are 26 tags in the tagset. Tags is the national language of India tagset (IL). The algorithm has a high success rate. Gujarati texts have a 92% accuracy rate. A total of 10,000 words make up the training corpus. The corpus for the test consists of 5,000 words. Researchers have trained a CRF on

Gujarati, which provides an insight into the language. With a 92% accuracy rate, researchers have learned from their experiments. It was discovered that if language-specific rules could be formed into CRF features, the accuracy would improve and can be attained at extremely high levels. The CRF is a group of people who work together to learn from both labeled (600 sentences) and untagged (600 sentences) data. In addition, there are untagged data, which total 5,000 sentences. We can deduct from the errors that the training was ineffective. As the amount of data grows, the number of terms that are unknown decreases will be found in the corpus of the exam, which will Increase the precision. We might also make use of a machine, lexicons, morphs, and other usable resources whenever they are constructed, etc. [19].

In 2013, Miral Patel and Prem Balani proposed a study on Clustering Algorithm for the Gujarati language. They assert that natural language processing is still a work in progress. Natural language processing entails assessing the language's structure and then accurately tagging each word with its grammar basis. Researchers have a set of 50,000 tagged words and are attempting to cluster those Gujarati words using a provided technique; they have also defined their own processing algorithm. There are numerous clustering algorithms available, for example, single linkage, complete linkage, and average linkage. The number of clusters that will be produced is unknown; therefore, it all depends on the data set provided. Clustering is a step in the stemming process. Stemming is the process of extracting the root from its term. "Cats", for example, is equal to cat+s, which means cat is a noun and is not plural. After clustering the tagging words, the results were quite evident, although clustering may be done with multiple languages and tag sets. The conclusion is that clustering with tagged words gives more than 98% accuracy, and clustering on tagged words is an intermediate step on the way to stemming [20].

In 2017, Utkarsh Kapadia and Apurva Desai proposed work on Rule-Based Gujarati Morphological Analyzer. In this chapter, the author also proposes a Rule-Based Gujarati Morphological Analyzer. Gujarati, like other Indo-Aryan languages such as Hindi and Marathi, has a large morphological vocabulary. Many NLP applications, such as machine translation, grammar inference, and information retrieval, require morphological analysis. Researchers have provided a morphological analyzer based on a rule-based approach in this research. We created a Gujarati morphological analyzer based on hand-coded inflectional and derivational rules derived from word and affix regularities. For sentence tagging, the same method was employed in the point-of-sale (POS) tagger. In addition, we have included a full account of the morphological phenomena found in

Gujarati. A lexical dictionary of root words is constructed. Manually generated rules are created in collaboration with linguists. The analyzer program accepts a Gujarati sentence and outputs its grammar class, gender, number, tense, and personal information, as well as its root words. Both inflectional and derivational morphemes are supported by the tool. Using material from essays and short stories, we were able to achieve an accuracy of 87.48%. Aside from its high accuracy, rule-based systems have some drawbacks. For starters, creating an exhaustive rule-based system for any natural language is tough. Second, if a suffix does not match any of the rules, the system may not produce any results. It is difficult to develop rules independent of language because they are so reliant on it [21].

The Gujarati Language: Research Issues, Resources, and Proposed Method on Word Sense Disambiguation is a paper that Tarjni Vyas and Amit Ganatra proposed in 2019. In this publication, the researchers spoke about the Gujarati language, word sense disambiguation (WSD), and Gujarati Wordnet. Because of this, it has been discovered that the deep learning strategy performs better in Gujarati WSD. However, one of its drawbacks is that it requires a large number of information sources, without which preparation is almost impossible. On the other hand, it employs information sources to choose word meanings in a particular context. Because of this, deep learning techniques seem to be more suitable for handling word meaning disambiguation, yet the procedure can never be simple since real languages are ambiguous. The contextual similarity between an example and an unknown paragraph is compared using this model. Use a growing number of training examples and choose the optimal features that include most of the vocabulary for effective generalization. Compared to Hindi, Gujarati has a more advanced use of colloquial terminology, which also differs by area. The Gujarati alphabet may contain words with multiple spellings since it is a more complex alphabet than the English alphabet. It is also likely that some data were saved with misspelled words since the writer of the data did not speak Gujarati. While solutions are suggested here, they are only the beginning; most strategies have not yet been investigated in Gujarati [22].

In 2019, Lata Gohil and Dharmendra Patel suggested a project on sentiment analysis of film reviews in the Gujarati language using machine learning. This text discusses the authors' description of how language is handled in a typical manner. It controls the presentation of data to determine the source's intention for the material. The cause might be attributed to either gratitude (positive) or academic investigation (negative). This work presents a connection between the results achieved by implementing the calculation arrangement using different classifiers, such as K-nearest-neighbor (KNN) and multinomial naive Bayes (MNB). The collected data have been carefully evaluated by considering the extreme film datasets and their correlation with the available evidence. This study examines the impact of the word level count vectorizer and term frequency-inverse document frequency (TF-IDF) on the sentiment analysis of films. The investigators found that the MNB classifier produces more precise outcomes when utilizing the TF-IDF vectorizer compared to the count vectorizer. Additionally, the KNN classifier yields similar accuracy results for both the TF-IDF and the count vectorizer. The researchers in this study compiled a film review dataset in the Gujarati language. They then applied two distinct machine learning-based classification techniques to this dataset, using count and TF-IDF vectorizer components. The objective was to evaluate the sentiment of movie evaluations in the Gujarati language. The findings of the sentiment study indicate that the TFIDF vectorizer features provide superior outcomes in comparison to the count vectorizer features. After evaluating the performance parameters of accuracy, recall, precision, and F-score, we found that the MNB model provided more accurate predictions when using TF-IDF data compared to count vectorizer data [23].

In 2020, Saif Ali Alsaidi, Ahmed T. Sadiq, and Hasanen S. Abdullah presented a study on the classification of English poetry utilizing text mining and rough set theory. The contributors discussed text classification, a technique of classifying articles into predetermined groups based on their textual content. This work aims to address the issue of categorizing poems in English into specific categories using text-mining techniques and machine-learning algorithms. In the suggested model, the authors used text preprocessing on the document file to decrease the number of features and minimize dimension. The preprocessing step transforms the textual poem into features and eliminates irrelevant features using text mining techniques such as tokenization, stop word removal, and stemming. To further reduce the feature vector, two feature selection methods were employed. The categorization task was then performed using rough set theory as the machine learning algorithm. The authors achieved a successful classification rate of 88% with the proposed model. Categorizing English poems is challenging without performing preprocessing steps to reduce the number of features. One way to achieve this is by removing stop words, which eliminates unimportant features. However, even after this step, there may still be a large number of features. To further reduce the features, stemming can be applied to bring many words back to the same root word. Finally, a filter for feature selection can be used to select the most significant characteristics [24].

In 2021, Bhavin Mehta and Bhargav Rajyagor proposed work on Gujarati poetry classification based on emotions using deep learning. They attempted to express emotions through Gujarati poetry by utilizing a range of characteristics found in Gujarati poems. Gujarati poems provide a unique perspective on sentiment capturing in this study. “Kavan” is a collection of Gujarati poems based on the Indian concept expressed in “Navarasa”. More than 300 poems were grouped into nine feelings indicated in “Navarasa” in the anthology. The result accuracy of the emotion classification task from the Gujarati poetry corpus was determined to be up to 87.62%. The authors discovered that there was limited research study for the classification of Gujarati poetry on the basis of the notion of “Navarasa” after conducting a literature review. The emotions expressed in the poem have a close link to the poem; poets have written poems that are full of emotions. The authors of this study experimented with the Indian idea of “Navarasa” to establish an automatic system for classification of poetries based on emotions present in the poem. The authors have compiled a collection of poems known as the “Kavan” Gujarati poetry corpus. They found that their system functioned admirably for emotion identification and categorization from Gujarati poetry based on the results obtained [25].

In this research work, we presented a proposed model for word segmentation or “sandhi” in Gujarati language using a purely engineering-based approach. A slight effort is made to create a link between the concept of Gujarati Grammar and the reality of its execution. As a result, not every South Asian language is familiar with the grammar notion Sandhi. However, Gujarati, Tamil, Sanskrit, Devnagri, and Hindi all have a significant role in this concept. The term “language” in NLP refers to natural languages such as Gujarati, Hindi, and English, which humans use to communicate on a daily basis. The majority of NLP research has focused on English and other European languages. In the last few years, NLP research on Indian languages such as Gujarati has begun. This paper’s main focus is on demonstrating the road map for implementing the Gujarati grammatical term “sandhi ().” Sandhi is a word segmentation technique that is found in most South Asian languages, including Devnagri, Sanskrit, Hindi, and Gujarati, as well as Chinese and Thai. “Sandhi causes phonetic alteration at the word boundaries of a written chunk (small section), and the sounds at the conclusion of the word combine to form a single chunk of the character sequence.” The rule-based application of “sandhi” is the focus of our attention. Gujarati language (grammar) has its unique set of rules for mixing consonants, vowels, and modifiers, just like every other Indian scripting language. We have come up with a set of guidelines for putting "sandhi ()" into practice. There are numerous sandhi rules, each denoting a distinct set of phonetic alterations described in Gujarati grammatical tradition. The Sandhi does not modify the meaning or grammatical structure of the words in question. Sandhi is a voluntary procedure that is solely dependent on the writer's awareness. They presented a method for Gujarati sandhi splitting based on raw text words. Though the approach is only described for Gujarati, it might be used for any language if some stated rules are followed [26].

In 2022, Parita Shah, Priya Swaminarayan, and Maitri Patel stated in their work, "A Sentiment Analysis of Gujarati Text using Gujarati Senti word Net", that sentiment analysis is a critical component of decision-making. This region sees a lot of research for the English language. In comparison to English, there is very little development in this subject for Indian languages. Gujarati is a nearly undiscovered language for this assignment. People prefer to use their local language on the Internet; therefore, more data in the form of movie reviews, product evaluations, and social media posts, among others, are available in regional languages, necessitating the need to mine these data in order to comprehend their opinions. Several tools and resources have been produced for the English language, but just a handful have been established for Indian languages. Gujarati is a language with limited resources for this task. The goal of this study is to create a sentiment lexical resource for Gujarati that can be used to analyze sentiment in Gujarati literature. Gujarati SentiWordNet is based on Hindi SentiWordNet and IndoWordNet's synonym relations of words. We make a twofold contribution: the Gujarati SentiWordNet is being created, while the Gujarati corpus is being generated in order to assess the lexical resource that has been created. The usefulness of the generated resource is demonstrated by the evaluation result. By utilizing synonym relations, this method is used to construct G-SWN using Hindi SentiWordNet and IndoWordNet. The resulting resource can be used to analyze Gujarati text for sentiment. The proposed method can be used to build sentiment lexical resources for any of the IWN languages. The Gujarati tweets corpus was created to test the lexical resource that was generated. Two annotators annotated the corpus for positive and negative polarity classes. Cohen's kappa, a statistical measure of inter-annotator agreement, was 0.55 for this corpus. The resulting annotated corpus contains 863 tweets, 442 of which are positive and 421 are negative. The accuracy of G-SWN utilizing this gold standard corpora was 52.72% for unigram presence and 52.95% for simple scoring classifiers. Results demonstrate the moderate performance of G-SWN. G-SWN provides the baseline for further studies. Future work can become comprehensive by the creation and use of antonym family members. By incorporating WSD, higher accuracy can be achieved [27] as given in Algorithm 5.1: Ghazal Chand Identifier.

Proposed Algorithm:

Algorithm 5.1: Ghazal Chand Identifier

```
Input: Ghazal Line
Output: Chand of Given Line for Ghazal
Initialize: an array array[] with words of ghazal , token string

1.  array[]=split[poem]; // delimiter = ‘ ‘(space)
2.  for(word:array)
3.  for (characterpoint : word)
4.  stack.Push(character point)
5.  while(stack!=empty)
6.  if(size==1) then tokenstring+=’1’ and pop()
7.  Else
8.  Top = stack.pop()
9.  Below = stack.pop()
10. If(‘top’+’below’ = la )then tokenstring+=’1’
11. Else tokenstring+=’2’
12. Initialize Map<Integer, String> lagastring , fill with pre-
    defined tokens
13. For (Map.Entry<Integer,String> var : lagastring.entrySet())
14. Comparator percentage (tokenstring,var)
```

This is an algorithm of Ghazal Chand Identifier; it helps us to identify the Chand of ghazal. Initialize the map of Chand that is meant to match a line inputted for Ghazal Chand. Split the line with space as delimiter. Iterate over each word and then fill it in stack. Get the top two values and compare them to its combination. If it is a single letter, then add “1” to the token string. If the combination of code points corresponds to the letter લા, then add “1” to the token; if the combination of code points corresponds to the letter ગા, then add “2” to the token . Similarly, generate the token and pass it to a percentage calculator parser function whose parameters are the values of the already initialized Chand tokens. Obtain the matching percentage from this function as shown in Figure 5.6.

5.5 Proposed System

There are five phases in the approach we use to create ghazal apps. Before working on the first line of any ghazal, you may want to read the first step (i/p લીમડાને આવી ગયો તાવ જીરે). The second phase is a transitional stage where the ghazal lines are combined into a single string without any white space in between. (short after trimming: લીમડાનેઆવીગયોતાવજીરે). In the previous step, the matras of the string produce tokens of numbers 1 for લ and 2 for ગા (Generation of token: 212222122122). In the following stage, the converted string of tokens 1 and 2 is compared to the system's predefined Chands. If the dataset matches, the Chand is created as an output as presented in Figures 5.7 to 5.13, respectively.

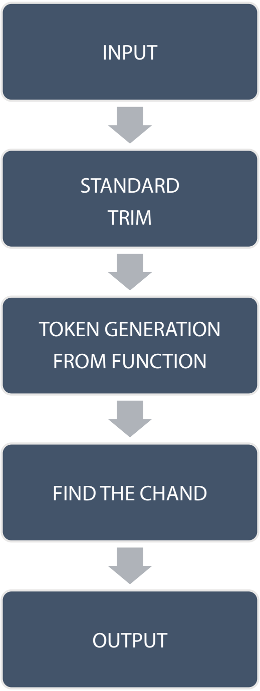


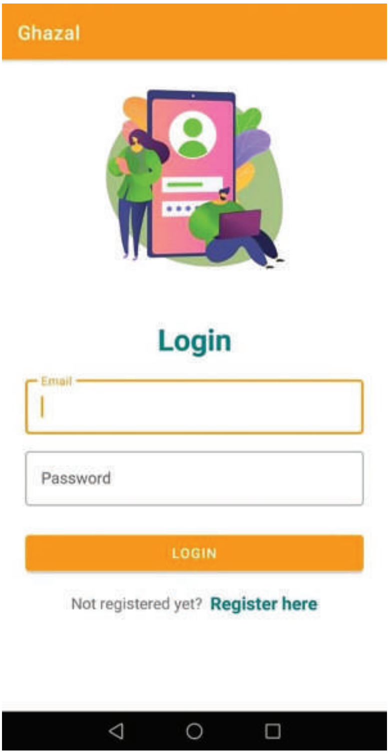
Figure 5.6 Proposed system.

GHAZAL

ગુજરાતી

ગુજરાતી ભાષાની પ્રથમ ગઝલ
છંદ તપાસતી એપ્લિકેશન

Figure 5.7 Splash screen.



Ghazal

Login

Email

Password

Not registered yet? Register here

Figure 5.8 Login screen.

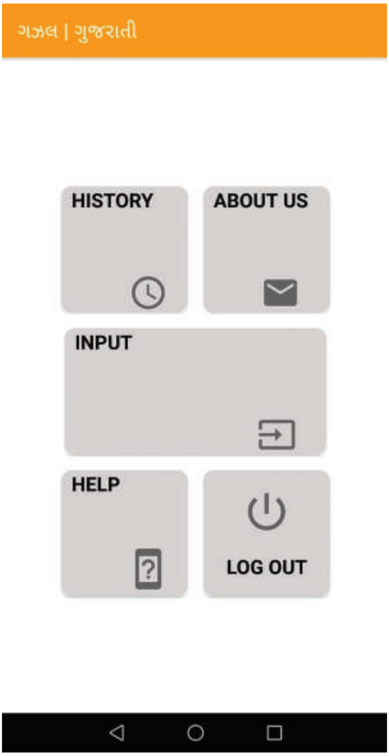
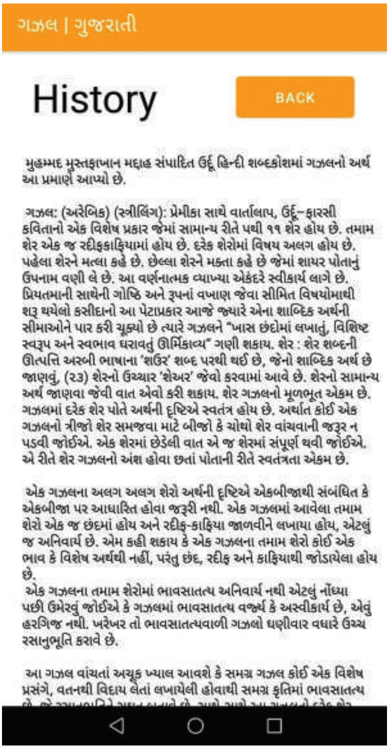


Figure 5.9 Home screen.



મુહમ્મદ મુસ્તફાખાન મદ્દાહ સંપાદિત ઉર્દૂ હિન્દી શબ્દકોશમાં ગઝલનો અર્થ આ પ્રમાણે આપ્યો છે.

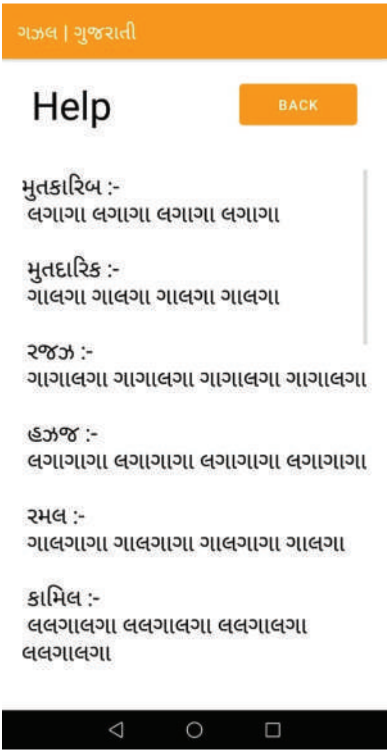
ગઝલ: (અરેબિક) (સ્ત્રીલિંગ): પ્રેમીકા સાથે વાર્તાલાપ, ઉર્દૂ–ફારસી કવિતાનો એક વિશેષ પ્રકાર જેમાં સામાન્ય રીતે ૫થી ૧૧ શેર હોય છે. તમામ શેર એક જ રદીફકાફિયામાં હોય છે. દરેક શેરોમાં વિષય અલગ હોય છે. પહેલા શેરને મત્લા કહે છે. છેલ્લા શેરને મક્તા કહે છે જેમાં શાયર પોતાનું ઉપનામ વણી લે છે. આ વર્ણનાત્મક વ્યાખ્યા એકંદરે સ્વીકાર્ય લાગે છે. પ્રિયતમાની સાથેની ગોષ્ઠિ અને રૂપનાં વખાણ જેવા સીમિત વિષયોમાથી શરૂ થયેલો કસીદાનો આ પેટાપ્રકાર આજે જ્યારે એના શાબ્દિક અર્થની સીમાઓને પાર કરી ચૂક્યો છે ત્યારે ગઝલને "ખાસ છંદોમાં લખાતું, વિશિષ્ટ સ્વરૂપ અને સ્વભાવ ધરાવતું ઊર્મિકાવ્ય" ગણી શકાય. શેર : શેર શબ્દની ઊત્પત્તિ અરબી ભાષાના 'શઉર' શબ્દ પરથી થઈ છે, જેનો શાબ્દિક અર્થ છે જાણવું, (૨૩) શેરનો ઉચ્ચાર 'શેઅર' જેવો કરવામાં આવે છે. શેરનો સામાન્ય અર્થ જાણવા જેવી વાત એવો કરી શકાય. શેર ગઝલનો મૂળભૂત એકમ છે. ગઝલમાં દરેક શેર પોતે અર્થની દૃષ્ટિએ સ્વતંત્ર હોય છે. અર્થાત કોઈ એક ગઝલનો ત્રીજો શેર સમજવા માટે બીજો કે ચોથો શેર વાંચવાની જરૂર ન પડવી જોઈએ. એક શેરમાં છેડેલી વાત એ જ શેરમાં સંપૂર્ણ થવી જોઈએ. એ રીતે શેર ગઝલનો અંશ હોવા છતાં પોતાની રીતે સ્વતંત્રતા એકમ છે.

એક ગઝલના અલગ અલગ શેરો અર્થની દૃષ્ટિએ એકબીજાથી સંબંધિત કે એકબીજા પર આધારિત હોવા જરૂરી નથી. એક ગઝલમાં આવેલા તમામ શેરો એક જ છંદમાં હોય અને રદીફ-કાફિયા જાળવીને લખાયા હોય, એટલું જ અનિવાર્ય છે. એમ કહી શકાય કે એક ગઝલના તમામ શેરો કોઈ એક ભાવ કે વિશેષ અર્થથી નહીં, પરંતુ છંદ, રદીફ અને કાફિયાથી જોડાયેલા હોય છે.

એક ગઝલના તમામ શેરોમાં ભાવસાતત્ય અનિવાર્ય નથી એટલું નોંધ્યા પછી ઉમેરવું જોઈએ કે ગઝલમાં ભાવસાતત્ય વર્જ્ય કે અસ્વીકાર્ય છે, એવું હરગિજ નથી. ખરેખર તો ભાવસાતત્યવાળી ગઝલો ઘણીવાર વધારે ઉચ્ચ રસાનુભૂતિ કરાવે છે.

આ ગઝલ વાંચતાં અચૂક ખ્યાલ આવશે કે સમગ્ર ગઝલ કોઈ એક વિશેષ પ્રસંગે, વતનથી વિદાય લેતાં લખાયેલી હોવાથી સમગ્ર કૃતિમાં ભાવસાતત્ય

Figure 5.10 History screen.



Help

મુતકારિબ :-
લગાગા લગાગા લગાગા લગાગા

મુતદારિક :-
ગાલગા ગાલગા ગાલગા ગાલગા

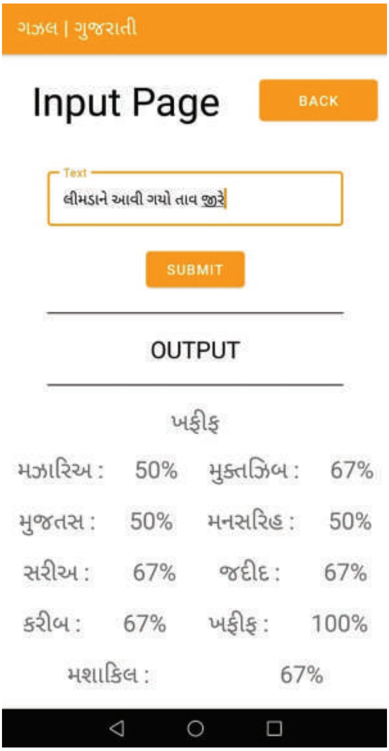
રજઝ :-
ગાગાલગા ગાગાલગા ગાગાલગા ગાગાલગા

હઝજ :-
લગાગાગા લગાગાગા લગાગાગા લગાગાગા

રમલ :-
ગાલગાગા ગાલગાગા ગાલગાગા ગાલગા

કામિલ :-
લલગાલગા લલગાલગા લલગાલગા
લલગાલગા

Figure 5.11 Help screen. Type of chanda.



ગઝલ | ગુજરાતી

Input Page

BACK

Text: લીમડાને આવી ગયો તાવ જીરે

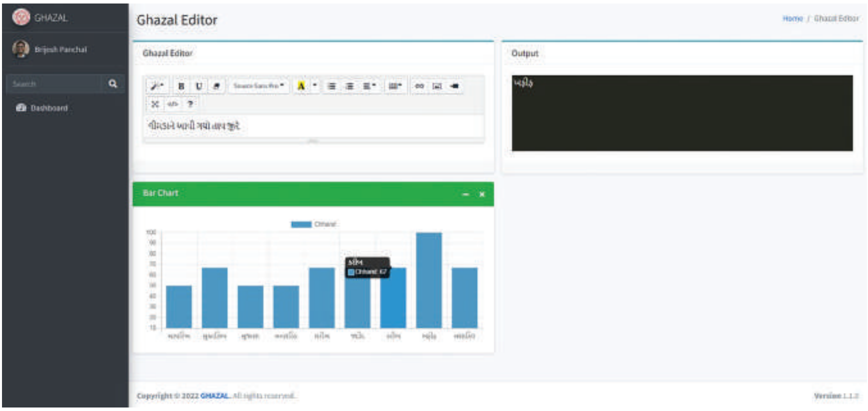
SUBMIT

OUTPUT

ખફીફ

મઝારિઅ :	50%	મુક્તઝિબ :	67%
મુજતસ :	50%	મનસરિહ :	50%
સરીઅ :	67%	જદીદ :	67%
કરીબ :	67%	ખફીફ :	100%
મશાકિલ :	67%		

Figure 5.12 Ouput for Khafif Ghazal.



GHAZAL

Brijesh Panchal

Search

Dashboard

Ghazal Editor

Home / Ghazal Editor

Ghazal Editor

લીમડાને આવી ગયો તાવ જીરે

Output

ખફીફ

Bar Chart

Copyright © 2022 GHAZAL. All rights reserved.

Version 1.1.2

Figure 5.13 Website output for Khafif Ghazal.

5.6 Conclusion

This system is an initial or first attempt in the field of the Gujarati NLP research community. One can identify the Gujarati Ghazal Chanda through this system. Among the many popular genres of poems in Indian literature is the Ghazal. In this paper, researchers of this chapter developed a cross-platform to identify the Ghazal meter. A total of 19 meters are used in this system. Simple Python programming has been used in this proposed system. With the help of this system, budding writers of Gujarati may improve their poetry writing. Gujarati language history, grammar, and steps of the Gujarati Ghazal writing process are described in this chapter. Overall, this chapter's content may be quite beneficial to Gujarati language lovers.

References

1. Acharya, S., *Kacchi sabdávali [Kutchi vocabulary)*, Gujarat Vidyapith, Ahmedabad, 1966.
2. Acharya, S., *Halari dialect*, Gujarat Vidyapith, Ahmedabad, 1985.
3. Bhandari, A., *Gujarati vakyaracanã [Gujarati syntax)*, University Grantha Nirmāṇa Board, Ahmedabad, 1990.

   Bhayani, H., *Gujarati bhāṣhanum aitihäsika vyakarana [A historical grammar of Gujarati)*, Gujarat sahitya akādami, Gandhinagar, 1988.
4. Cardona, G., *A Gujarati reference grammar*, University of Pennsylvania Press, Philadelphia, 1965.
5. Cardona, G. and Suthar, B., Gujarati, in: *The Indo-Aryan languages*, G. Cardona and D. Jain (Eds.), pp. 659–697, Routledge, London, 2003.
6. Dave, J., *Colloquial Gujarati: a complete language course*, Routledge, London, 1995.
7. https://www.omniglot.com/writing/gujarati.htm
8. Dave, T., *A study of the Gujarati language in the 16th century*, Royal Asiatic Society, London, 1935.
9. Deshpande, P., *Universal English-Gujarati dictionary*, Oxford University Press, Bombay, 2022.
10. Dwyer, R., *Gujarati: a complete course for beginners*, Hodder Headline PLC, London, 1995.
11. Gajendragadkar, S., *Parsi-Gujarati: a descriptive analysis*, University of Bombay, Bombay, 1974.
12. Joshi, U., Raval, A., Shukal, Y. (Eds.), Gujarati sahityano itihasa, in: *History of Gujarati literature*, vol. 4, Gujarātī sahitya parisad, Ahmedabad, 1973-1981.

13. Vishal, History of Ghazal, in: *Something about Everything*, 31 May 2016, Web.
14. Mishra, A. K., Tyagi, A. K., Dananjayan, S., Rajavat, A., Rawat, H., Rawat, A., Revolutionizing government operations: The impact of artificial intelligence in public administration. in: *Conversational Artificial Intelligence*, pp. 607–634, 2024.
15. Snyder, B.H. and Margery, Poetry: What Is the History and Definition of a Ghazal?, in: *Thought Co*, 22 Apr. 2016, Web.6.
16. Doty, G., A Short History of the Ghazal, in: *A Short History of the Ghazal*, 2007, Web.
17. Nahar, S., Pithawa, D., Bhardwaj, V., Rawat, R., Rawat, A., Pachlasiya, K., Quantum technology for military applications. in: *Quantum Computing in Cybersecurity*, pp. 313–334, 2023.
18. Sikarwar, R., Shakya, H. K., Kumar, A., Rawat, A., Advanced security solutions for conversational AI. in: *Conversational Artificial Intelligence*, pp. 287–301, 2024.
19. Patel, C. and Gali, K., Part-Of-Speech Tagging for Gujarati Using Conditional Random Fields. *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, January 2008, pp. 117–122.
20. Patel, M. and Balani, P., Clustering Algorithm for Gujarati Language. *Int. J. Sci. Res. Dev.*, 1, 3, 2321–0613, 2013, ISSN (online).
21. Pithawa, D., Nahar, S., Bhardwaj, V., Rawat, R., Dronawat, R., Rawat, A. Quantum computing technological design along with its dark side. in: *Quantum Computing in Cybersecurity*, pp. 295–312, 2023.
22. Vyas, T. and Ganatra, A., Gujarati Language: Research Issues, Resources and Proposed Method on Word Sense Disambiguation. *Int. J. Recent Technol. Eng. (IJRTE)*, 8, 2S11, September 2019, ISSN: 2277-3878.
23. Namdev, A., Patni, D., Dhaliwal, B. K., Parihar, S., Telang, S., Rawat, A., Potential threats and ethical risks of quantum computing. in: *Quantum Computing in Cybersecurity*, pp. 335-352, 2023.
24. Alsaidi, S.A., Sadiq, A.T., Abdullah, H.S., English poems categorization using text mining and rough set theory. *Bull. Electr. Eng. Inf.*, 9, 4, 1701–1710, August 2020, DOI: 10.11591/eei.v9i4.1898, ISSN: 2302-9285
25. Mehta, B. and Rajyagor, B., Gujarati Poetry Plassification based on Emotions using Deep Learning. *Int. J. Eng. Appl. Sci. Technol.*, 6, 1, 358–362, 2021, ISSN No. 2455-2143.
26. Noonia, A., Beg, R., Patidar, A., Bawaskar, B., Sharma, S., Rawat, H., Chatbot vs Intelligent Virtual Assistance (IVA). in: *Conversational Artificial Intelligence*, pp. 655-673, 2024.
27. Shah, P., Swaminarayan, P., Patel, M., Sentiment analysis on film review in Gujarati language using machine learning. *Int. J. Electr. Comput. Eng. (IJECE)*, 12, 1, 1030–1039, February 2022, DOI: 10.11591/ijece.v12i1. PP-1030-1039, ISSN: 2088-8708.

6

Cancer Classification and Detection Using Machine Learning Techniques

Syed Jahangir Badashah[1]*, Afaque Alam[2], Malik Jawarneh[3], Tejashree Tejpal Moharekar[4], Venkatesan Hariram[5], Galiveeti Poornima[6] and Ashish Jain[7]

[1]Department-ECE, Sreenidhi Institute of Science and Technology, Yamnampet, Hyderabad, Telangana, India
[2]Department of Computer Science and Engineering, Bakhtiyarpur College of Engineering, Patna, Bihar, India
[3]Oman College of Management and Technology, Muscat, Oman, INTI International University, Subang, Malaysia
[4]Yashwantrao Chavan School of Rural Development, Shivaji University, Kolhapur, Maharashtra, India
[5]Department of Practice of Medicine, Vinayaka Mission's Homoeopathic Medical College and Hospital, A Constituent College of Vinayaka Mission's Research Foundation Deemed to be University, Salem, Tamil Nadu, India
[6]Presidency University, Bangalore, India
[7]Department of Computer Applications, The Bhopal School of Social Sciences, Bhopal, India

Abstract

A doctor may use pictures from an MRI, CT, or X-ray to find bone cancer. The manual process is laborious and time-consuming, requiring specific understanding. Therefore, developing a system that can automatically discern between healthy and cancerous bones is crucial. Cancer-damaged bones feel different than neighboring healthy bones. Nevertheless, a few of the sample's images have morphological characteristics associated with both healthy and malignant bone. It becomes difficult to categorize them as a result. This article presents Machine learning and image processing based statistical pattern recognition framework for classification and detection of bone cancer. MRI and CT scan images are used as input data set

**Corresponding author*: syd_jahangir@yahoo.co.in

Rajesh Kumar Chakrawarti, Ranjana Sikarwar, Sanjaya Kumar Sarangi, Samson Arun Raj Albert Raj, Shweta Gupta, Krishnan Sakthidasan Sankaran and Romil Rawat (eds.) Natural Language Processing for Software Engineering, (95–112) © 2025 Scrivener Publishing LLC

in this framework. Images quality is improved using the Gaussian Elimination. In the end, classification is performed using the CNN and LSTM. Models are trained and tested with sufficient data.

***Keywords*:** Cancer, classification, detection, gaussian elimination, machine learning, accuracy, CNN, LSTM

6.1 Introduction

Early detection is vital in all types of cancer, but particularly in breast cancer, due to its chronic nature and the usual delay in diagnosis. Given the current situation, it is imperative to take immediate action on the problem of predictive detection. The considerable global research undertaken on cancer detection encompasses several aspects, allowing for speculation on the potential results of a predictive technique. To reduce the negative impact of cancer on health, it is crucial to tackle issues related to accurately evaluating the risk for those with a high susceptibility and facilitating cancer prognosis for patients. Predictive cancer detection models not only identify individuals with a high risk of developing cancer, but they also provide current knowledge on how to assess the probability of cancer. Consequently, it simplifies the process of arranging and carrying out clinical cancer trials, so facilitating the advancement and evaluation of risk-benefit indices. This, in turn, enables an evaluation of the financial burden and effect of cancer. However, the previously indicated models may be used to evaluate the effectiveness and delivery of the medicines [1]. A malignant tumor, also known as a malignant neoplasm, is an aberrant proliferation of cells that has the ability to metastasize to other parts of the body, resulting in the disease known as cancer. Conversely, not all tumors are cancerous, hence benign tumors do not metastasize. Several symptoms that are often seen in various cancers include: a palpable mass, abnormal or excessive hemorrhaging, a chronic cough, significant unintentional weight loss (>10%), irregular bowel movements, and so forth. Over one hundred different types of cancer may affect a person. Cancer is characterized by uncontrolled cell proliferation, which occurs when the normal regulation of cell growth and production is disrupted [2]. According to the statistics, tobacco is accountable for around 22% of all cancer-related deaths. Alcohol use, sedentary lifestyle, obesity, and bad dietary habits together account for 10% of all deaths. Infections, ionizing radiation, and environmental pollutants are among the least common of these dangers. Although these factors may play a part in cellular gene modifications, it is conceivable that the formation of cancer necessitates considerably more substantial changes in DNA. Approximately 5-10% of cancer occurrences are attributed to genetic defects that are passed down through generations. While cancer signs may be apparent, other diagnostic procedures, such as a biopsy or medical imaging, may be necessary to definitively establish the diagnosis. Unfortunately, even being diagnosed and undergoing several rounds of treatment, there is still a silent risk of the malignant form reappearing, which requires precise identification and prompt prediction. The prediction models developed using machine learning theory will be essential after all these factors have been considered. The field of cancer prediction has widely used machine learning techniques. The potential applications of predictive analytics in the healthcare business are many. Due to the crucial significance of identifying cancer at an early stage and accurately forecasting its progression, the utilization of real-world data on cancer patients has immense potential for the advancement of more efficient prediction techniques. However, to create accurate prediction models, it is essential to include appropriate and relevant features [3].

Cancer is becoming more prevalent and lethal, with a rising number of cases and fatalities worldwide. Conversely, there is a perception that the increasing amount of data might provide new opportunities for tackling this issue and creating a predictive solution based on data analysis. There seems to be a significant association between verified cases and medical records, which, with further investigation, may uncover factors that assist in predicting cancer. Therefore, it is essential to use methodologies that rely on data analysis and machine learning techniques [4].

6.2 Machine Learning Techniques

Prior to embarking on an in-depth exploration of machine learning methods and their practical implementation in particular situations, it is crucial to fully comprehend the meaning and importance of machine learning. Machine learning is a specific area of research within the science of artificial intelligence. It involves the use of probabilistic, statistical, and optimization techniques to learn from past examples. The earlier training provides advantages for the classification of new data, pattern identification, and trend prediction. Data analysis and interpretation are performed using machine learning methodologies, similar to those used in statistics. On the other hand, machine learning methods that rely on statistics use unconventional optimization strategies that utilize model data or patterns, along with absolute conditionality (IF, THEN, ELSE), conditional probabilities (the probability of X given Y), and Boolean logic (AND, OR, NOT). All of the aforementioned strategies closely resemble the way humans really assess and categorize objects [5].

Machine learning has always relied on established methods of probability and statistics for detection. Machine learning, in contrast, is far more powerful and practical since it enables the generation of suggestions or conclusions that traditional methods cannot achieve. Statistical approach often includes the use of multivariate regression and correlation analysis. Effective conventional techniques depend on the assumption of variables' independence and linear organization for the interpretation of data. Traditional statistical approaches often prove inadequate when variables exhibit conditional dependence and their interactions are nonlinear. Machine learning excels in the field of cancer diagnosis. The majority of genetic structures exhibit nonlinearity, with their parameters being time-dependent. There are many simple physical formations that have a linear shape and whose characteristics are essentially unrelated to each other. There is no guarantee that cancer diagnoses with machine learning will provide satisfactory outcomes. A comprehensive understanding of the issue and the constraints of the available data are essential for the success of any method. Furthermore, a key factor in the effectiveness of the approach is a thorough understanding of the assumptions and limitations of the algorithms. A machine learning experiment may be deemed successful by meticulous planning, appropriate use of learners, and verification of outcomes. The traditional saying suggests that if low-quality data are inputted, the output will also be of low quality. Likewise, if the quantity of variables surpasses the anticipated instances, an unnecessary succession of learners will be created. The selection of input data does not impact the amount of execution for these sets of learning algorithms [6]. The "curse of dimensionality" refers to the difficulty of handling a large number of variables within a limited number of situations. This "curse" also impacts several statistical procedures and machine learning techniques. To resolve this problem, one may either decrease the number of variables (features) or increase the number of training samples. A section-to-feature ratio higher than 5:1 is often expected. The breadth and variety of the training set are crucial. It is crucial that the training examples include a diverse array of data types that the learner is expected to come across. Overtraining or overfitting happens when there is a scarcity of examples and little diversity in the training data. When confronted with fresh information, an excessively trained learner may have difficulty in understanding or categorizing it. Traditional statistics have been shown to be sometimes more effective and accurate than machine learning. This suggests that the user's original hypothesis about the data's interdependence and lack of linearity was inaccurate. This does not imply a deficiency in machine learning; instead, it depends on making the correct choice. Uniform treatment is not universally observed in all machine learning approaches [7].

Diverse tactics provide disparate outcomes on various challenges. For instance, although some machine learning algorithms effectively evaluate the size of genetic domains, others perform quite poorly. Furthermore, certain methodologies may not be appropriate for the current situation owing to underlying assumptions or data prerequisites. Identifying the optimal strategy for a certain problem is a challenging task. Therefore, it is recommended to use several machine learning techniques to conduct experiments on each provided training dataset. There is a common misunderstanding about machine learning that stems from the fact that its ability to recognize patterns and trends often results in unexpected findings. Nevertheless, by meticulous analysis of the data, a human specialist may identify various trends and patterns. Machine learning simplifies the process of sickness detection by providing a less complicated method for identifying patterns and creating classification systems [8–12].

The desired result is the foundation for classifying these algorithms. Supervised learning approaches depend on a "prescient provider," often a teacher, who supplies a labeled set of training data or examples. The tagged samples assist the program in acquiring knowledge on how to optimize the input data for efficient production planning. An instance of this may be a categorized training dataset that has a series of indistinct images depicting the numeral 8. Under the supervision of an instructor, an apprentice may practice the skill of taking notes on all the images labeled as "8." The goal is to provide an accurate and undistorted representation of the method's outcomes. The classroom is the most prevalent setting for pupils to acquire knowledge. Unsupervised learning involves a set of examples that lack labels. Whether a novice is able to see the prototype or discern the categories mostly depends on their own observation skills. Graduate students are the individuals who are most prone to demonstrating this particular type of learning. Unsupervised learning algorithms include methods such as K-means clustering, hierarchical clustering, and self-organizing feature maps (SOMs). These techniques produce clusters from unprocessed, unlabeled, or unclassified raw data. In the future, these clusters might potentially be used to begin classifiers or provide categorization suggestions [13–16].

The SOM approach demonstrates a distinct design of an artificial neural network. The technique relies on altered weights that correlate to input vectors, using a training set that includes a grid of artificial neurons. Initially, the SOM approach is intended to serve as a substitute for the functional characteristics of the genetic brain. The procedure starts with a collection of artificial neurons, each of which has a substantial function in the ultimate outcome. In a winner-take-all process, which is a competitive network, they function as a component where a vector of inputs is designated as the leading candidate and the weights are modified to closely resemble the input vector. This method is dependent on this particular facet of competitive learning. Each one is surrounded by a cluster of nodules. When a node emerges victorious in a conflict, the weights of its neighboring nodes are also altered, but not to the same extent. A lesser displacement in weight happens when the neighbor is geographically far from the winner. A vast quantity of cycles is thus used in doing this operation repeatedly for each input vector. The outcome of the competition might significantly differ based on the inputs provided. The end result is a neural network that can effectively connect certain clusters or prototypes in the input dataset to its output nodes. It is crucial to bear in mind that all machine learning algorithms, both before and after a cancer diagnosis, depend on supervised learning. Moreover, conditional prospects or conditional judgments often serve as the foundation for the categorization and organization procedures of these supervised learning systems.

The use of Support Vector Machines (SVMs) for cancer diagnosis is a modern machine learning technique. Support vector machines (SVMs) are well recognized in the field of machine learning, but they have not yet gained much recognition in the fields of pre-diagnosis and cancer diagnostics. For instance, in the context of breast cancer, SVMs may be used indirectly to distinguish between people with favorable and unfavorable prognoses by comparing the number of axillary metastases to the tumor mass. Two separate clusters have been found. The objective of the support vector machine (SVM) machine learning system is to identify the linear equation that optimally optimizes the distance between the two groups. If a single cluster were to include additional variables like as volume, metastases, and estrogen receptor satisfaction, the rows of division would be transformed into planes. If more variables were provided, a hyperplane would be used to explain the division. Subsequently, the support vectors, which delineate the boundaries between the two classes, reinforce the position of this hyperplane. In the SVM method, the data is first divided into two groups using the maximum range to produce a hyperplane. Given this circumstance, it implies that the region between the hyperplane and the closest examples (the boundary) is maximized. Support vector machines (SVMs) use non-linear kernels to accomplish nonlinear classifications. Using a non-linear kernel may help convert data from a linear characteristic space to a non-linear element space numerically. Applying diverse kernels to various data sets automatically enhances the presentation of an SVM classification. SVMs, similar to ANNs, may be used for diagnosing many difficulties including prototype recognition and classification, medical diagnosis, verbal and content identification, protein utility calculation, and deviating from manual handwriting inspection [17–19].

6.3 Review of Machine Learning for Cancer Detection

The collaboration between techniques for identifying new information and ways for safely processing current data is essential. Implementing this will enhance the usability of a cancer prediction algorithm, perhaps facilitating timely identification and effective management. The number 20 is represented by the numeral [20]. Various renowned global research institutions generate and distribute publicly available cancer datasets that are considered benchmarks in the area. The UCI repository and the SEER database are widely used in several studies, making them the preferred choices among researchers. Reference [21] used the Wisconsin breast cancer dataset to construct and assess many prediction models. To assess the effectiveness of different data mining methods, the authors in [22] used the SEER dataset. An effective method for using machine learning theory is to formally express the problem of cancer prediction as a binary classification algorithm. This method is effective for the majority of cancer types. To diagnose an event or assess its severity, the mathematical model will analyze the connections and patterns among the necessary information and predict the classification of the current stage. Support Vector Machines (SVMs) are categorization algorithms that have been used in recent breast cancer research [23, 24]. The utilization of an SVM-based classifier model enables accurate classification of cancer-related data, regardless of whether it is in numerical or visual format. This powerful technique [43–46], based on the theory of optimal separating hyperplanes, facilitates early detection of cancer and prediction of survival outcomes for patients in advanced stages. The research on breast cancer presented in reference [25] demonstrated that Support Vector Machines (SVMs) were the best suitable method. In a study conducted by [26], an SVM model trained on clinical data achieved a remarkable accuracy of 95.5% in predicting breast cancer. The clinical prognosis of new patients was accurately predicted by using the medical history of patients with cancer who had similar characteristics. This approach had the highest accuracy rate of 66.7% when evaluated with a linear kernel support vector machine [27]. A crucial prognostic indicator for the long-term survival of patients with gastric cancer was used to train a support vector machine using 5-fold cross-validation [28]. The classifier has an average sensitivity of 88.5% and an average specificity of 78.5%. Support vector machines have been used in recent studies to develop classification techniques. According to reference [29], support vector machines (SVMs) are the optimal choice for using breast cancer data in the field of predictive analytics. The Wisconsin breast cancer database, consisting of 699 tuples, was analyzed using predictive analytics. One of the four approaches outlined in [30], the support vector machine (SVM), was used for this analysis. In [31], it was shown that a support vector machine (SVM) was the most efficient method for selecting features in many breast cancer datasets. The authors of [32] achieved greater prediction accuracies in identifying cancer gene markers by using a supervised transductive SVM technique. DTs, or decision trees, are characterized by their ability to quickly acquire and carry out prediction tasks. These predictors are based on information entropy. Reports indicate that the use of DT resulted in a 93% accuracy rate for predicting outcomes in the SEER database [22]. To differentiate between normal and malignant stomach tissues, the researchers used the Classification and Regression Tree (CART) approach to classify their Raman infrared spectra [33]. The predictive sensitivity for these input parameters was determined to be 88.9% on the validation dataset, while the specificity was reported to be 92.9%. The most effective model for forecasting the five-year survival rates of lung cancer patients, using a combined database from Michigan and Harvard, was found to be a Decision Tree (DT) based on the C5.0 algorithm, as stated in reference [34]. Cancer prediction research heavily relies on the use of decision tree (DT) based classifiers. The breast cancer datasets in [31] were examined using decision trees (DTs) and classification and regression trees (CARTs), both with and without feature selection. To predict susceptibility to breast cancer using the standard Wisconsin dataset, a Decision Tree (DT) was used as one of the six classification algorithms indicated in reference [35]. A cross validated technique using a Decision Tree (DT) based on the J48 algorithm obtained an accuracy of 97% in diagnosing non-small cell lung cancer and its two primary subtypes [36]. Artificial Neural Networks (ANNs) are becoming more prevalent in cancer prediction [37]. One of the five unique prediction methods employed in [21] was a multilayer neural network. An artificial neural network (ANN) was trained utilizing several characteristics, including age and mammography findings, to achieve a high accuracy rate of 96.5% in predicting breast cancer susceptibility [38]. Using a testing data set consisting of 40 women in the control group and 40 women in the cancer group, a feed-forward artificial neural network (ANN) obtained prediction metrics of 82.5% specificity and sensitivity [39]. The purpose was to find bio-proteins that function as multiple indications for the identification of breast cancer. The use of Artificial Neural Networks (ANNs) in predicting cancer recurrence was extensively explained in reference [40]. The estimation and forecasting of survival for patients with non-small lung cancer was conducted using Artificial Neural Networks (ANNs) [41]. To forecast risk variables for the prognosis of gastric cancer patients, a back propagation artificial neural network (ANN) achieved an average classification rate of 84.16%. This was accomplished by using several pathological and laboratory feature [42]. K-Nearest Neighbors (KNN) is a fundamental approach of classification. This lazy-learning strategy is used to develop a classification rule, which relies on the occurrences in the training data. Its user-friendly interface makes it very popular among academics worldwide. Aloraini used the K-Nearest Neighbors (KNN) algorithm in their prediction methodology, as documented in reference [21] of the UCI Machine Learning (ML) repository. Instance-based classifiers, such as K-Nearest Neighbors, are often used while constructing cancer prediction models. Based on a study conducted in [29], which used the UCI machine learning library to predict breast cancer recurrence, it was shown that a KNN classifier was the most efficient method overall. One of the techniques used in the study cited in [35] also uses KNN.

6.4 Methods

This section presents a convolutional neural network based framework for cancer image classification and detection. MRI images are first preprocessed using Gaussian Filter to remove noises. Then images are classified by CNN and LSTM deep learning techniques as shown in Figure 6.1: A Framework for Cancer Image Classification and Detection.

To get the desired result, the picture must first be filtered and then enhanced. The results of segmentation performed on photographs obtained with a mobile phone may be affected by a broad variety of factors due to the nature of the photographs themselves. Resizing a picture, lowering the amount of noise in it, and improving it are all examples of processes that are included in pre-processing. It's possible for a digital picture to include a broad array of sounds. Because of this, there is a possibility of picture noise, which renders conventional thresholding ineffective. It is necessary to reduce the amount of noise in the photos. Image noise is the accidental change in a photograph's brightness or coloring that occurs across the image. A picture may include several kinds of noise, including Gaussian noise, noise with salt and pepper patterns, noise with shot patterns, and

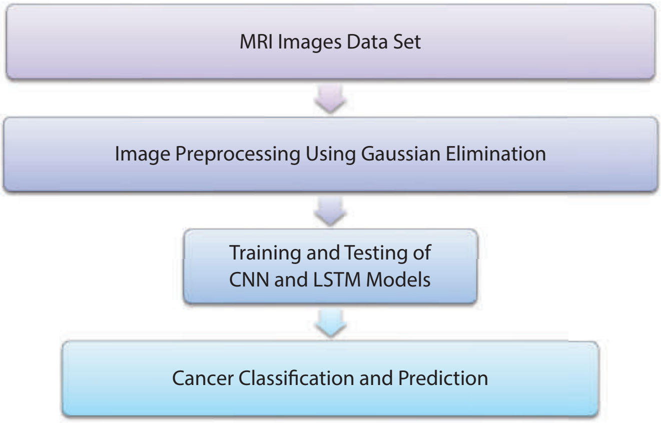

MRI Images Data Set
Image Preprocessing Using Gaussian Elimination
Training and Testing of CNN and LSTM Models
Cancer Classification and Prediction

Figure 6.1 A framework for cancer image classification and detection.

quantization noise. There is a possibility that filters such as the median and the Wiener can eliminate these blips in the data. A number of different morphological techniques may be used to reduce the overall loudness of an audio source. The median and Gaussian filters each have their own unique effect on how bright individual pixels are made to seem. To achieve the aim of reducing the amount of background noise, GF was used. In the process of Gaussian filtering, the significance of the intensity of any particular pixel is replaced by a weighted average of the intensities of the pixels that are next to it [18].

Because it is an RNN network with gates such as I/P, forget, and O/P in addition to an additional memory cell, the LSTM (Long Short-Term Memory) [25] algorithm is a well-liked option. The need for gradient descent may be avoided by using LSTM networks because of their capacity to remember information for a protracted period of time. This enables the networks to better recognize patterns and sequences. The data is kept current by using input and output gates in between the time steps. When using Equation 6.2, the information that will be stored in the memory cell is determined by the input gate I. The Forget Gate (f) erases all prior memories by reorganizing the cell's present state in accordance with Equation 6.1. This renders all previous information superfluous. The output gate (y), which is the last phase, is responsible for controlling the data that is sent on to the succeeding stage. As a result of the fact that all three gates are connected to the memory cell, it is possible to monitor the timing of the output. LSTM is shown in Figure 6.2: LSTM network.

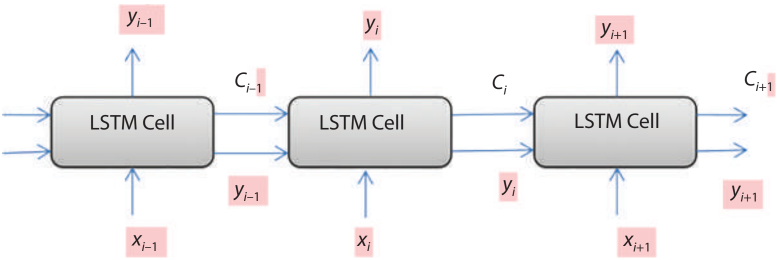


Figure 6.2 LSTM network.

$$f_t = \sigma(w_f.[y_{t-1}, x_t] + b) \tag{6.1}$$

$$i_t = \sigma(w_i.[y_{t-1}, x_t] + b) \tag{6.2}$$

The LSTM network is outlined in Figure 6.4. At each time step, an input i.e. the embedding x_i is fed into the network and the output y_i is determined based on the current embedding x_i, previous output y_{i-1} and the past cell state c_{i-1}. Cell state is capable of adding or removing the information.

There are several hidden layers in a CNN, and two of those levels are the convolution and pooling layers [26]. Image processing is one of its strong suits, and it has the ability to discover dependencies. The features are extracted from the information that is fed into the system using the convolution layer. Convolutional operations are carried out on the embedding matrix using this process. The embedding matrix is where the word embedding vectors that were generated by the word embedding approaches may be found stored. Following the convolution layer in a CNN is a layer called the pooling layer, which is responsible for performing operations such as dimensionality reduction and feature selection through pooling. It is conceivable to carry out a process that involves maximum pooling, a procedure that involves minimum pooling, or an operation that involves average pooling.

After the features have been acquired, they are utilized as input into a neural network that has already had all of its connections fully constructed. The activation functions are responsible for producing the output. The convolutional neural network is seen here in Figure 6.3: Convolution Neural Network. This study made use of two convolution layers together with average pooling to achieve its results.

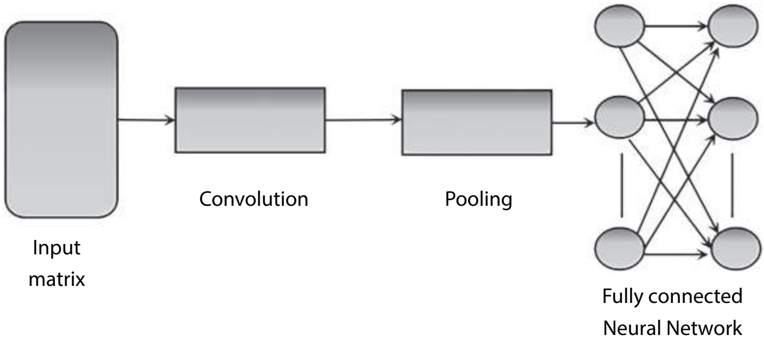
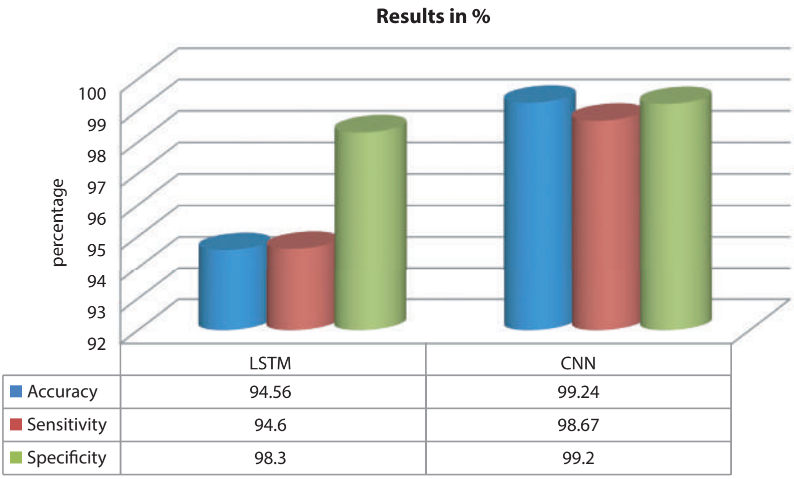


Figure 6.3 Convolution neural network.

6.5 Result Analysis

For experimental work, a total of 200 images were selected at random. 160 images were used for training of model, 20 images were used for validation and remaining 20 images were used for the testing of the model. Images were preprocessed to remove noise using Gaussian filter. Classification



	LSTM	CNN
🟦 Accuracy	94.56	99.24
🟥 Sensitivity	94.6	98.67
🟩 Specificity	98.3	99.2

Figure 6.4 Result comparison of classifiers.

was performed using CNN and LSTM deep learning techniques. Results are shown below in Figure 6.4: Result Comparison of Classifiers.

Sensitivity:

$$\text{Sensitivity} = \frac{\text{TP}}{(\text{TP} + \text{FN})}$$

Where TP stands for True Positive and FN stands for False Negative

Specificity:

$$\text{Specificity} = \frac{\text{TN}}{(\text{TN} + \text{FP})}$$

Where TN stands for True Negative and FP stands for False Positive

Accuracy

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{(\text{TN} + \text{TP} + \text{FN} + \text{FP})}$$

6.6 Conclusion

Early detection is essential in cases of breast cancer, primarily because late diagnosis is prevalent. The matter of anticipatory detection is currently of utmost significance. The existing literature has not provided enough evidence to apply the prediction technique universally in response to these issues. To effectively manage cancer-related illness and death, it is crucial to have precise cancer risk assessment for persons at high risk and to define the prognosis for cancer patients. Predictive cancer models have enabled recent advancements in risk assessment and the identification of high-risk patients. Developing such predictive models facilitates the process of strategizing and designing clinical cancer studies, as well as constructing benefit-risk indices. These models can assist in estimating the cost, impact, and effectiveness of cancer treatments and management.

References

1. Yershova, K., Yuan, J.-M., Wang, R., Valentin, L., Watson, C., Gao, Y.-T., Hecht, S.S., Stepanov, I., Tobaccospecific n-nitrosamines and polycyclic aromatic hydrocarbons in cigarettes smoked by the participants of the shanghai cohort study. *Int. J. Cancer*, 139, 6, 1261–1269, 2016.
2. Louie, A.V., Haasbeek, C.J., Mokhles, S., Predicting overall survival after stereotactic ablative radiation therapy in early stage lung cancer: Development and external validation of the amsterdam prognostic model. *Int. J. Radiat. Oncol. Biol. Phys.*, 93, 1, 82–90, 2015.
3. Euhus, D.M., Smith, K.C., Robinson, L., Stucky, A., Olopade, O., II, Cummings, S., Garber, J.E., Chittenden, A., Mills, G.B., Rieger, P., Esserman, L., Crawford, B., Hughes, K.S., Roche, C.A., Ganz, P.A., Seldon, J., Fabian, C.J., Klemp, J., Tomlinson, G., Pretest prediction of BRCA1 or BRCA2 mutation by risk counselors and the computer model brcaproj. *Natl. Cancer Inst.*, 94, 844–851, 2002.
4. Tyrer, J., Duffy, S.W., Cuzick, J., A breast cancer prediction model incorporating familial and personal risk factors. *Statist. Med.*, 23, 1111–1130, 2004.
5. Bianchi, F., Bracci, R., Rosati, S., Galizia, E., Belvederesi, L., Loretelli, C., Giorgetti, G., Giorgi, F., Cellerino, R., CRCAPRO: A statistical model to evaluate the risk of mmr mutations, *J. Clin. Oncol.* in:*, 2005 ASCO Annual Meeting Proceedings*, June 2005, vol. 23, p. 9693.
6. Colditz, G.A., Rosner, B.A., Speizer, F.E., Risk factors for breast cancer according to family history of breast cancer. *J. Natl. Cancer Inst.*, 88, 6, 365–371, 1996.
7. Spasic̀, I., Livsey, J., Keanec, J.A., Nenadic̀, G., Text mining of cancer-related information: Review of current status and future directions. *Int. J. Med. Inf.*, 83, 9, 605–623, June 2014.
8. Brown, W.R. and Ahnen, D.J., The international health care burden of cancers of the gastrointestinal tract and liver. *Cancer Res. Front.*, 1, 1, 1–9, February 2015.
9. Mishra, A. K., Tyagi, A. K., Dananjayan, S., Rajavat, A., Rawat, H., Rawat, A., Revolutionizing Government Operations: The Impact of Artificial Intelligence in Public Administration. in: *Conversational Artificial Intelligence*, pp. 607–634, 2024
10. Bellazzi, R. and Zupan, B., Predictive data mining in clinical medicine: Current issues and guidelines. *Int. J. Med. Inf.*, 77, 2, 81–97, February 2008.
11. Ramachandran, P., Girija, N., Bhuvaneswari, T., Early detection and prevention of cancer using data mining techniques. *Int. J. Comput. Appl.*, 97, 13, 48–53, July 2014.
12. Matsuno, R.K., Costantino, J.P., Ziegler, R.G., Anderson, G.L., Li, H., Pee, D., Gail, M.H., Projecting individualized absolute invasive breast cancer risk in asian and pacific islander american women. *J. Natl. Cancer Inst.*, 103, 12, 951–961, June 2011.

13. Chlebowski, R.T., Chen, Z., Anderson, G.L., Rohan, T., Aragaki, A., Lane, D., Dolan, N.C., Paskett, E.D., McTiernan, A., Hubbell, F.A., Adams-Campbell, L.L., Prentice, R., Ethnicity and breast cancer: Factors influencing differences in incidence and outcome. *J. Natl. Cancer Inst.*, 97, 6, 439–448, March 2005.
14. Spitz, M.R., Hong, W.K., Amos, C., II, Wu, X., Schabath, M.B., Dong, Q., Shete, S., Etzel, C.J., A risk model for prediction of lung cancer. *J. Natl. Cancer Inst.*, 99, 9, 715–726, March 2007.
15. Spitz, M.R., Etzel, C.J., Dong, Q., Amos, C., II, Wei, Q., Wu, X., Hong, W.K., An expanded risk prediction model for lung cancer. *Cancer Prev. Res.*, 1, 4, 250–254, 2008.
16. Hoggart, C. *et al.*, A risk model for lung cancer incidence. *Cancer Prev. Res.*, 5, 6, 834–846, 2012.
17. Nahar, S., Pithawa, D., Bhardwaj, V., Rawat, R., Rawat, A., Pachlasiya, K., Quantum technology for military applications. in: *Quantum Computing in Cybersecurity*, pp. 313–334, 2023.
18. Suthar, H., Rawat, H., Gayathri, M., Chidambarathanu, K., Techno-Nationalism and Techno-Globalization: A Perspective from the National Security Act. in: *Quantum Computing in Cybersecurity*, 137–164, 2023.
19. Amir, E., Evaluation of breast cancer risk assessment packages in the family history evaluation and screening programme. *J. Med. Genet.*, 40, 11, 807–814, 2003.
20. Rawat, R., and Rajavat, A., Perceptual Operating Systems for the Trade Associations of Cyber Criminals to Scrutinize Hazardous Content. *IJCWT*, 14, 1, 1–19, 2024.
21. Sikarwar, R., Shakya, H. K., Kumar, A., Rawat, A., Advanced Security Solutions for Conversational AI. in: *Conversational Artificial Intelligence*, 287–301, 2024.
22. Delen, D., Walker, G., Kadam, A., Predicting breast cancer survivability: a comparison of three data mining methods. *Artif. Intell. Med.*, 02, 34, 113–127, 2005.
23. Chauhan, D., Singh, C., Rawat, R., Dhawan, M., Evaluating the Performance of Conversational AI Tools: A Comparative Analysis. in: *Conversational Artificial Intelligence*, 385–409, 2024.
24. Tseng, C.-J., Lu, C.-J., Chang, C.-C., Chen, G.-D., Application of machine learning to predict the recurrence-proneness for cervical cancer. *Neural Comput. Appl.*, 24, 13116, 2014.
25. Kim, W., Kim, K.S., Lee, J.E., Noh, D.Y., Kim, S.-W., Jung, Y.S., Development of novel breast cancer recurrence prediction model using support vector machine. *J. Breast Cancer*, 2, 15, 230–238, 2012.
26. Eshlaghy, A.T., Poorebrahimi, A., Ebrahimi, M., Razavi, A.R., Ahmad, L.G., Using three machine learning techniques for predicting breast cancer recurrence. *J. Health Med. Inform.*, 4, 124, 2013.
27. Chan, L.W.C., Machine learning of patient similarity: A case study on predicting survival in cancer patient after locoregional chemotherapy, in: *IEEE*

International Conference on Bioinformatics and Biomedicine Workshops (BIBMW), 2010.

28. Pithawa, D., Nahar, S., Bhardwaj, V., Rawat, R., Dronawat, R., Rawat, A., Quantum Computing Technological Design Along with Its Dark Side. in: *Quantum Computing in Cybersecurity*, 295–312, 2023.
29. Rawat, R., Telang, S., William, P., Kaur, U., CU, O. K. (Eds.). *Dark Web Pattern Recognition and Crime Analysis Using Machine Intelligence*. IGI Global, 2022.
30. Rathi, M. and Gupta, C., An approach to predict breast cancer and drug suggestion using machine learning techniques. *ACEEE Int. J. Inf. Technol.*, 1, 4, 23–31, 2014.
31. Lavanya, D. and Usha Rani, K., Ensemble decision tree classifier for breast cancer data. *Int. J. Inf. Technol. Convergence Serv.*, 1, 2, 17–24, 2012.
32. Maulik, U., Mukhopadhyay, A., Chakraborty, D., Geneexpression based cancer subtypes prediction through feature selection and transductive svm. *IEEE Trans. Biomed. Eng.*, 60, 4, 1111–1117, 2013.
33. Namdev, A., Patni, D., Dhaliwal, B. K., Parihar, S., Telang, S., Rawat, A., Potential Threats and Ethical Risks of Quantum Computing. in: *Quantum Computing in Cybersecurity*, 335–352, 2023.
34. Mihalache, L., Suresh, S., Yaosuo, X., Manjrekar, M., Modeling of a small distribution grid with intermittent energy resources using MATLAB/SIMULINK, in: *Proc. 2011 IEEE Power and Energy Society General Meeting*, pp. 1–8, 2011.
35. Ravi kumar, G., Ramachandra, G.A., Nagaman, K., An efficient prediction of breast cancer data using data mining techniques. *Int. J. Innov. Eng. Technol.*, 2, 4, 139–144, August 2013.
36. Venkat, D.M., Rasheed, M.A., Ali, M., Classification of lung cancer subtypes by data mining technique, in: *IEEE International Conference on Control, Instrumentation, Energy and Communication (CIEC)*, 2014.
37. Shrivastava, M., Patil, R., Bhardwaj, V., Rawat, R., Telang, S., Rawat, A., Quantum Computing and Security Aspects of Attention-Based Visual Question Answering with Long Short-Term Memory. in: *Quantum Computing in Cybersecurity*, 395–412, 2023.
38. Noonia, A., Beg, R., Patidar, A., Bawaskar, B., Sharma, S., Rawat, H., Chatbot vs Intelligent Virtual Assistance (IVA). in: *Conversational Artificial Intelligence*, 655–673, 2024.
39. Zhang, F. and Chen, J.Y., A neural network approach to multibiomarker panel development based on LC/MS/MS proteomics profiles: A case study in breast cancer, in: *22nd IEEE International Symposium on Computer-Based Medical Systems*, vol. *2009*, 2009.
40. Ritthipravat, P., Artificial neural networks in cancer recurrence prediction, in: *IEEE Computer Engineering and Technology 2009. ICCET09*, vol. 02, 2009.

41. Lee, L.G., Lee, C.Y., Park, I.K., Kim, D.J., Park, S.Y., Kim, K.D., Chung, K.Y., Number of metastatic lymph nodes in resected non-small cell lung cancer predicts patient survival. *Pubmed/ NCBI*, 01, 85, 211–215, January 2008.
42. Rawat, R., Chakrawarti, R.K., Sarangi, S.K., Choudhary, R., Gadwal, A.S., Bhardwaj, V., eds. *Robotic Process Automation*. John Wiley & Sons, 2023.
43. Shrivastava, M., Patil, R., Bhardwaj, V., Rawat, R., Telang, S., Rawat, A., Quantum Computing and Security Aspects of Attention-Based Visual Question Answering with Long Short-Term Memory, in: *Quantum Computing in Cybersecurity*, pp. 395–412, 2023.
44. Rawat, R., Chakrawarti, R.K., Raj, A.S.A., Mani, G., Chidambarathanu, K., Bhardwaj, R., Association rule learning for threat analysis using traffic analysis and packet filtering approach. *Int. J. Inf. Technol.*, *15*, 6, 3245–3255, 2023.
45. Rathi, M. and Rajavat, A., Investigations and Design of Privacy-Preserving Data Mining Technique for Secure Data Publishing. *Int. J. Intell. Syst. Appl. Eng.*, *11*, 9s, 351–367, 2023.
46. Dubey, P. and Rajavat, A., Effective K-means clustering algorithm for efficient data mining, in: *2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN)*, IEEE, pp. 1–6, 2023, May.

7

Text Mining Techniques and Natural Language Processing

Tzu-Chia Chen

Department of Artificial Intelligence, Tamkang University, New Taipei City, Taiwan

Abstract

Text mining is an important branch of data mining that is used to analyze the text data. Text data is any type of data like structured data, unstructured data, and semi-structured data. All types of data are collected from different sources such as multimedia applications, mobile apps, digital systems, etc. These data are beneficial to get a good insight, meaningful results. We use data mining techniques like Support Vector Machine (SVM), Random Forest (RF), Multilayer Perception (MLP), Naive Bayes (NB), etc. to analyze the hidden relationship between data. We use three kinds of data types in text mining.

***Keywords*:** Text mining, machine learning, natural language processing, SVM, random forest

7.1 Introduction

Text thematic organizing is an essential human activity in the modern digital era. The demand for strong algorithms capable of interpreting and extracting valuable insights from massive amounts of digital data has grown in tandem with the proliferation of high-tech devices and software. For this aim, data mining has shown to be an effective approach. To find useful patterns and correlations in massive datasets, data mining sometimes involves looking at the data from several angles and then summarizing it [1]. Data can take the form of any processable fact, picture, table, number, or text.

Email: tzuchiachen1688@gmail.com

Rajesh Kumar Chakrawarti, Ranjana Sikarwar, Sanjaya Kumar Sarangi, Samson Arun Raj Albert Raj, Shweta Gupta, Krishnan Sakthidasan Sankaran and Romil Rawat (eds.) Natural Language Processing for Software Engineering, (113–126) © 2025 Scrivener Publishing LLC

By systematically searching for and analyzing intriguing patterns in large amounts of unstructured material, text miners hope to unearth previously unknown and potentially valuable information [2]. Figure 7.1 shows the process of text mining, which is also known as intelligent text analysis or knowledge finding in text.

The raw dataset that is acquired is frequently inconsistent and ambiguous, as seen in Figure 7.1. This is the reason why the preprocessing stage is so important for applications that include text mining. The phases that are often included in the process of preparing unstructured data for text mining are visualized in Figure 7.2. The natural language processing techniques that are included in these methods include stemming, the removal of stop words, and part of speech tagging (POS), which assigns tags to each word, such as nouns, adjectives, and so on.

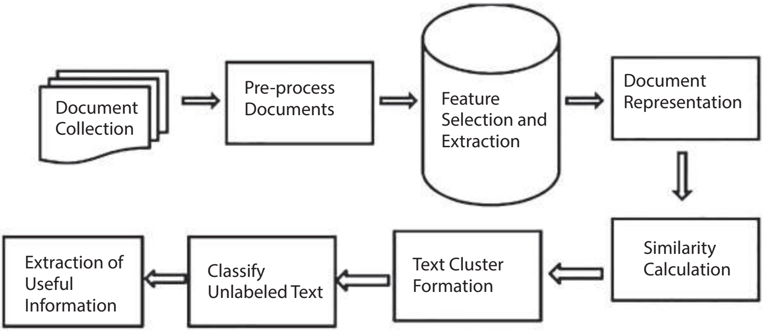


Figure 7.1 Steps in text mining.

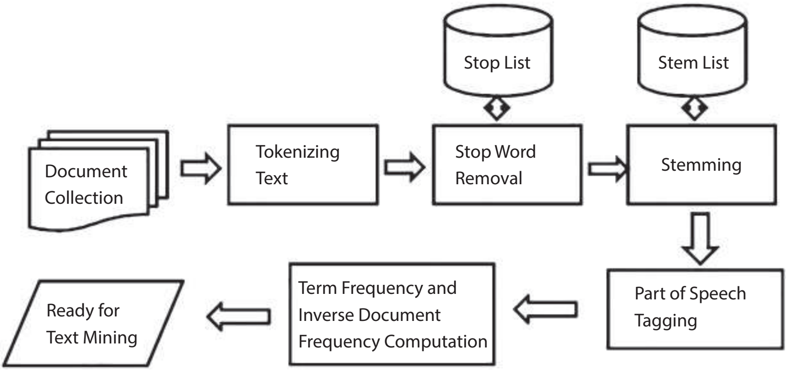


Figure 7.2 Stages of preprocessing text.

The vector space model, often known as the VSM, is the basis for the majority of text retrieval algorithms. This model represents data by considering texts to be a collection of words. To generate a feature vector of words for each text, the VSM makes use of either the term frequency or the term weight [3]. The objective of the feature selection process is to pick a subset of the characteristics of the original document that are capable of functioning effectively as a representation. To develop new features that are more pertinent to the problem at hand, the process of feature extraction entails changing the information that is being entered. To quantify the degree to which two documents are similar to one another, it is usual practice to utilize pair-wise similarity measures like as the Cosine and Jaccard similarity measures [4]. These measures are based on feature vectors. This is accomplished by employing similarity values that are bound by a threshold to build groups of connected documents. In the following step, further unlabeled material can be categorized using a variety of different methods as shown in Figure 7.2: Stages of Preprocessing Text.

7.2 Text Classification and Text Clustering

The goal of text clustering, an unsupervised method, is to arrange text documents into clusters based on their similarities and differences [5]. Image processing, pattern recognition, machine learning, and market segmentation are just a few of the many areas where clustering is now being studied. Clustering of text texts is the main emphasis of this thesis. When dealing with extremely huge datasets that contain several features of diverse sorts, text mining compounds the difficulties of clustering. As a result, applicable clustering methods face novel computational challenges. In response to these needs, a number of algorithms have surfaced and found use in practical data mining applications [6].

Most clustering techniques rely on two things: the availability of a similarity measure that may identify documents to be grouped together and a certain number of clusters. It is not always possible to predict in advance how many clusters will emerge. Documents are represented based on phrase or pair wise context, and the similarity link between the sentences may be found using tree representation similarity, according to a critical review of contemporary document clustering techniques [7]. But as the amount of data grows, the computational complexity and time required by this method skyrocket. Even when two words appear in a document at the same frequency, one of them may be more important to the meaning of the sentences than the other. This is a big problem with most text mining approaches since they use term frequency to determine a term's value in the document [8].

Data mining and machine learning communities devote a great deal of time and energy to studying classification. In classification, as opposed to clustering, the challenge is to assign labels to freshly encountered testing data; in contrast, clustering makes use of an existing set of pre-classified training data. To learn the descriptions of classes and to classify fresh data, the training data is utilized. Researchers from several fields have spent decades studying the classification problem. Text, multimedia, social networks, and even biological data are just a few of the many issue areas that might benefit from categorization. Additionally, the issue might arise in other contexts, including those with imprecise or flowing data. The data domain and issue scenario significantly impact the underlying algorithms, making classification a relatively broad topic. The test cases that cannot be observed are divided into categories according to the class label in the classification issue. Although clustering is often used to group instances, there is a significant distinction between the two issues. When it comes to clustering, the segmentation is accomplished by looking for similarities between the feature variables, without knowing how the groups are structured beforehand. Segmentation for classification purposes is based on a training dataset that stores information about the grouping structure as a target variable. Therefore, although clustering and data segmentation are both founded on ideas of similarity, in practice, it is possible to make substantial departures from similarity-based segmentation. Thus, supervised learning is used to describe the classification problem, whereas unsupervised learning is used to describe clustering. Some of the most common approaches to classification include decision tree induction, fuzzy logic, genetic algorithms, case-based reasoning, k-nearest neighbor classifier, and Bayesian networks.

7.3 Related Work

The movie review dataset is very useful for individuals to think about and make decisions about a wide variety of real-world challenges. A lot of people get other people's opinions before they make a final choice. Knowledge may be extracted from a wide range of study areas, including sentiment analysis, opinion mining, and text classifications. Reviewing films is one use case for sentiment analysis. Analyzing people's emotions isn't as simple as the term suggests. The categorization [18] of the movie review data from the dataset requires a number of preprocessing processes. The leading relevant research in sentiment analysis and text mining has been presented by researchers. The writers have gone over the sentimental content and the major effects of word of mouth on internet stores and customers. To determine the degree of reliance between different polarity ranges of terms, they used a coefficient of correlation (CC) and a natural language processing (NLP) method for positive/negative bag of words (BOW) [9]. Various viewpoints published in Roman-Urdu and English have been culled from a blog, and the authors have addressed text categorization in a WEKA setting. In terms of accuracy, Naïve Bayes (NB) surpasses Decision Tree (DT) and K-Nearest Neighbors (KNN), two of the three classification approaches they have used. The authors have covered topics such as deep learning-based sentiment analysis of Twitter databases. To analyze Twitter data, they used a Convolution Neural Network (CNN), which is great for extracting data from bigger texts and, by extension, for sentiment analysis. When compared to SVM and NB approaches, the proposed CNN method performs better. Studies on regional languages, such as Kannada (the official language of Karnataka), have focused on sentiment analysis. Using both ML and semantic methods, they analyzed sentiment and classified texts from Kannada online articles into an English-language dataset; ML methods outperformed semantic learning methods on average with the Kannada dataset. The authors have suggested an approach to user nature identification that relies on SoftMax based attitude detection. By considering the user's attitude, the accuracy and reliability of emotion prediction may be significantly enhanced. Utilizing tweets collected from the microblogging platform Twitter, the proposed algorithm is assessed [10]. With the use of sentiment analysis, the authors were able to improve their findings while analyzing text-based data from Chinese microblogs. The authors propose a novel approach to sentiment analysis for text categorization using text-based hidden Markov models (Text Mms). By including text semantic similarity, the authors' suggested tenor factorization approach enhances the factor's primary purpose of discovering comparable users, revealing the data's underlying characteristics, and forecasting users' preferences. Customers' reviews have been worked on by the authors. To do sentiment analysis, we randomly selected reviews [11]. Subject models and other popular CSTM text analysis techniques are shown to be ineffective when dealing with brief texts, according to the authors. We found that our CSTM beat the state-of-the-art short text topic conventional model in experiments conducted on real-world datasets. Wang, To examine the polarity of sentiment for brief texts, the authors have presented Senti-related, a unique cross-domain sentiment classification technique based on SRI. When working with unlabeled data from the target domain, Senti-related uses SRI to extend feature vectors. An issue that has sparked global discussion on social media, the Syrian refugee crisis has impacted millions of individuals and has prompted researchers to probe popular sentiment and opinion on the matter. They looked at pertinent tweets in both Turkish and English to see how people felt about the subject on Twitter. Additionally, they have examined the contrasting tone of the recovered tweets. To discover feature-opinion pairs and disclose the orientation of extracted opinions, the authors have presented two distinct approaches to building rule-based systems. The authors have proposed other classifiers for text data classification. Twenty newsgroups and IMDb have contributed to their data set [12]. Several methods, including KNN, RF, and NB, have been used by the writers to examine the movie evaluations. Opinion mining systems that express user's favorable or negative opinion at different levels have been the subject of research. We can extract feelings from the web and anticipate online client preferences using the same approach for predicting views. This might be useful for marketing research. Due to the importance of views, which we seek out anytime we must make a choice, a lot of recent study has focused on how our feelings and opinions are processed. Using a logistic regression classification model and an artificial bee colony algorithm, the authors have developed a new approach to spam detection. They have tested it on three separate datasets and shown that it can effectively handle high-dimensional data. Using text semantic analysis, the authors have proposed and implemented a system to better identify and categorize spam. Moreover, they have implemented spam detection in their respective domains using automatically collected semantic characteristics. Experts have deliberated how to identify and categorize ham data and e-mails that include both text and images. Three algorithms were used for spam e-mail classification: KNN, NB, and the reverse DB-SCAN method. The accuracy, precision, sensitivity, and specificity of these classifiers were found to be good when tested both before and after data preprocessing. Various evaluations pertaining to email spam filters have been proposed by the writers. A wealth of data on spam filters is included in this article [13]. To classify words for spam email filtering, researchers have proposed a feature selection-based semantic ontology. The spam email filter was built using Information Gain, Latent Dirichlet Allocation, a generative statistical model, and a semantic-based FST. Spam detection and detection-cum-analysis of hacked accounts are two interrelated issues that the authors have primarily addressed. up this publication, they have also filled up several gaps in the research that will be necessary for future studies. To identify spam, the authors have created a model that uses the Sender Policy Framework (SPF) protocol to examine the IP addresses of A and MX records. Using N-gram text data and an IDF feature selection approach, the authors have developed a new spam filter methodology. They have also provided an algorithm to balance distributions and a model for a regularized Deep Multilayer Perceptron Neural Network with rectified linear units called DBB-RDNNReL. Better text categorization accuracy was also attained when they evaluated spam filter performance with various ML techniques. The authors have researched the shortcomings of signature-based systems and spam blacklisting systems, and they have presented an ID3 algorithm for spam filtering based on the decision tree approach. According to [14], the suggested method outperformed competing algorithms in terms of accuracy. In the WEKA setting, the authors have implemented the Random Forest ML technique. Their spam email filter is quite effective and has very few features. To examine the efficacy of classifiers for spam and ham document categorization, the authors have used supervised ML algorithms such as NB, Perceptron, and C4.5. Naïve Bayes classifier, which outperforms other algorithms in terms of accuracy, has been proposed by them. The authors have classified spam emails using the SVM light program, which uses four kernel functions. Additionally, they have analyzed the dataset and computed several utility functions, such as TF, IDF, and TF-IDF. When applied to the email, the proposed SVM classifier improved the accuracy of the spam classification. When compared against more traditional methods of spam filtering, such as black lists and white lists, the authors' proposed Logistic Model Tree Induction approach in a WEKA environment outperformed the competition. When it comes to spam filtering, the authors have suggested using the Negative Selection Algorithm (NSA) to discover anomalies. For the Enron1 spam e-mail dataset, the suggested technique achieved an accuracy of 93.14 percent. To assess database performance, the writers have used distributed memory, distributed bag of words, cosine similarity, and auto encoder techniques. Due to the extensive representation of features, space complexity is a significant downside. Distributed memory and the distributed bag of words method have produced huge dictionaries for their respective word vectors, but the ever-increasing quantity of their vocabulary poses a dilemma when it comes to fixing the spam e-mail issue [15]. To lessen the spatial and temporal complexity, the authors propose using the Word Net ontology in conjunction with semantic based approaches and similarity metrics to reduce the number of textual characteristics that are extracted. By combining the FST with Principal Component Analysis (PCA) and Correlation Feature Selection (CFS), the authors were able to improve computing speed while simultaneously reducing spatial complexity. They have created a program called Concept Drift Analyzer that uses a k-fold cross-validation approach to accurately recognize ham and spam emails. When it comes to adversarial risk categorization, the authors have looked at the NB algorithm and the ACRA framework. An extension of anti-spam filtering, the authors have proposed the multi-objective optimization problem. This study used three new evolutionary multi-objective algorithms: SMS-EMOA, CH-EMOA, and MOEA/D, which are based on decomposition. For e-mail categorization with a low confidence level, our method outperforms another spam filter and optimizes the performance of heterogeneous ensemble classifiers. To train the classifier with high-quality data, the authors have suggested a novel method for producing phishing e-mail examples. Additionally, they have introduced quantitative assessment techniques, strengthened the generalizability of classifier control quantity sequence pairs, and built six resource generators and a communication connection selector. To overcome the uncertainty that arose from polysemy in spam e-mail detection, the authors propose using NB classification in conjunction with a conceptual and semantic similarity method. For permanent spam email categorization, the authors have presented a new model termed ELCADP. They evaluated the suggested model's performance with existing approaches for spam e-mail document categorization and found that it outperformed the competition. To categorize phishing emails, the authors have presented a new Remove Replacement Feature Selection Technique (RRFST) in addition to two decision tree methods [16]. The authors have put forward the ALO-Boosting technique for spam email classification. This approach makes use of ALO to choose the best feature subset, which is then fed into a boosting algorithm to improve classification accuracy. Irrelevant communications were a major factor in digital investigations, according to the authors. When it comes to digital investigations into spam emails, this letter gives a lot of useful information. The authors gathered information regarding spam e-mails and hint about hackers using five such cases. With the use of an e-mail spam corpus, the authors presented a text analytics approach that effectively detects threads and spam e-mails. To categorize spam, it used the text keyword matching approach with the corpus. This helped to avoid the inbox from being inundated with useless emails. To detect spam emails, the authors combined a logistic regression classification model with an artificial bee colony technique. To test their model's efficacy, they compared it to the existing model and utilized three separate datasets that were accessible to the public. Two studies have been conducted by the authors to detect SMS spam. The first trial assessed the suggested approach's performance, whereas the second employed a Bayesian network classifier testing method in conjunction with a cost-sensitive strategy [17].

7.4 Methodology

This section contains machine learning and feature selection-based methodology for text mining. This method consists of a text data set as input. First of all, features are selected using particle swarm optimization algorithm. Then machine learning algorithms are tested with testing data. The accuracy of algorithms is measured to correctly classify records. Finally, prediction of records is performed. The proposed methodology is shown below in Figure 7.3.

To improve its effectiveness, the PSO algorithm will take use of the fact that many different kinds of animals, including fish and birds, exhibit behavior similar to that of a swarm. Within the parameters of the search, each particle has a position and velocity that are completely unique to itself, and they are free to move in any way that they see fit. The velocity and mobility of the particle are still constrained as a direct consequence of

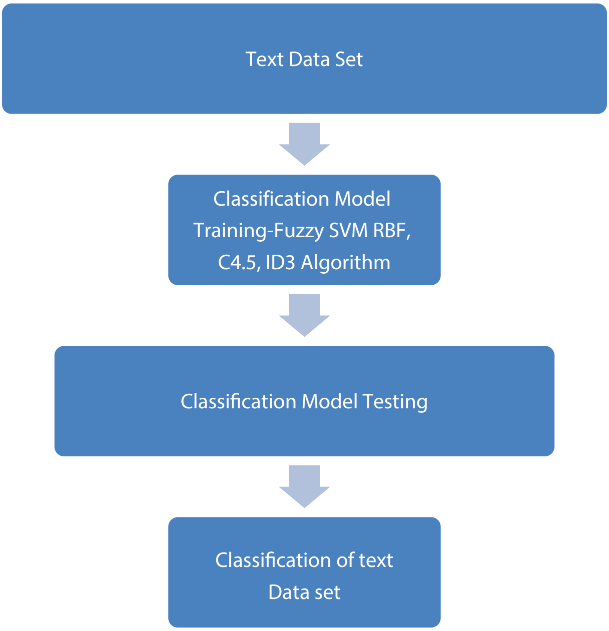

Text Data Set
Classification Model Training-Fuzzy SVM RBF, C4.5, ID3 Algorithm
Classification Model Testing
Classification of text Data set

Figure 7.3 Machine learning based framework for text mining.

this, and the particle is directed to the location of any previously successful particles. In addition to this, the particle is directed to the location of any other particles that have successfully completed their tasks in the past without causing any mishaps. It is now possible, as a result of the application of a predetermined set of rules to the technique in question, to characterize both the rate of motion and the location of each particle at every particular instant in time.

Identification 3 (ID3) is a method that use a decision tree to categorize instances in an iterative manner according to the values and features of those instances. It is generated from the most complex to the simplest using a collection of instances and the required features in that order. These groups were then further subdivided. When the set included inside a certain subtree is equal to the set of instances that belong to a comparable category, the set in question is designated as a leaf node within the given subtree. Information theoretical criteria are taken into consideration whenever a characteristic is being chosen for testing purposes. The reduction in entropy will result in an increase in the information gain at each node.

Decision trees [19, 20] are often used in machine learning models for the purpose of data classification [21–23] because of their speed and accuracy. When using this method, tree trimming may be accomplished in a number of different ways, making it a very adaptable practice. The result of the trimming is one that can be comprehended with ease. A number of academicians are of the opinion that taking down trees might be utilized as a means of overfitting. The C4.5 technique includes an iterative classification that is used to further improve the categorization of the data until only pure leaf nodes are left, at which point the process comes to a conclusion. This iterative classification can be found in the method's outline. By using this strategy, it is possible to get the most out of the training data, while also eliminating the demand for rules that only signal a specific behavior in the data. This is because it is practical to get the most out of the training data.

In recent years, the use of the technique known as the support vector machine (SVM) has seen significant growth in this sector. A helpful strategy for data categorization is the use of a linear or non-linear separation surface in the input space. This may be done either linearly or non-linearly. The support vector machine (SVM) seeks to construct a model that reliably predicts the goal values of new data by using just the characteristics of the test data. The data collected during the training phase are used to construct the model. The support vector machine (SVM) provides a dependable approach for pattern categorization that maintains a high level of accuracy while simultaneously lowering the danger of overfitting. When there are just two types of classes present in the data, SVM may be used. Using

SVM, data is categorized by locating the hyperplane that most effectively separates all of the data points into distinct categories. The hyperplane of a support vector machine (SVM) that has the greatest difference in margin between two classes of data is considered to be the optimal hyperplane for an SVM. When the slab is at its widest point parallel to the hyperplane, there are no data points located inside the boundary of the slab itself. Because of their close closeness to the hyperplane that separates the slab, the points along the slab's border are regarded to be support vectors. In fuzzy SVM, a membership value is assigned to each sample determined by its correlations with the values of other samples. There are many different ways in which each input sample contributes to the process that ultimately results in a choice being made. A radial basis function kernel has also been found to improve the performance of a support vector machine (SVM).

7.5 Conclusion

Text data is a treasure trove of information, but deciphering massive amounts of it requires a lot of work in terms of content organization, retrieval, and linking. The process of automatically extracting valuable information from massive amounts of unstructured text, such as blog posts, correspondence, internal reports (e.g., medical records), financial reports, research papers, account statements, and a plethora of web documents, among many others. Text mining is an emerging multidisciplinary discipline that brings together fields such as computational linguistics, statistics, data mining, machine learning, and information retrieval. Many businesses see text mining as a lucrative strategy because the vast majority of data is stored in text format (more than 80%). This chapter contains machine learning and feature selection-based methodology for text mining. This method consists of a text data set as input. First of all, features are selected using particle swarm optimization algorithm. Then machine learning algorithms are tested with testing data. The accuracy of algorithms is measured to correctly classify records. Finally, prediction of records is performed.

References

1. Mishra, A. K., Tyagi, A. K., Dananjayan, S., Rajavat, A., Rawat, H., Rawat, A., Revolutionizing Government Operations: The Impact of Artificial Intelligence in Public Administration. in: *Conversational Artificial Intelligence*, pp. 607–634, 2024.

2. Radovanovic, M. and Ivanovic, M., Text Mining: Approaches and Applications. *Novi Sad J. Math.*, 38, 3, 227–234, 2008.
3. Wang, X., Tang, J., Liu, H., Document clustering via matrix representation, in: *Proceedings of 11th IEEE International Conference on Data Mining*, pp. 804–813, 2011.
4. Schaeffer, E.S., Survey on Graph Clustering. Elsevier *Comput. Sci. Rev.*, 1, 1, 27–64, 2007.
5. Li, C.H., Yang, J.C., Park, S.C., Text categorization algorithms using semantic approaches, corpus-based thesaurus and WordNet. *Expert Syst. Appl.*, Elsevier, 39, 765–772, 2012.
6. Zhang, W., Yoshida, T., Tang, X., Wanga, Q., Text clustering using frequent itemsets. *Knowl.-Based Syst.*, Elsevier, 23, 379–388, 2010.
7. Lam, W. and Hwuang, R., An active learning framework for semi-supervised Document Clustering with language modeling. *Data Knowl. Eng.*, 68, 49–67, 2009.
8. Pithawa, D., Nahar, S., Bhardwaj, V., Rawat, R., Dronawat, R., Rawat, A., Quantum Computing Technological Design Along with Its Dark Side. in: *Quantum Computing in Cybersecurity*, pp. 295–312, 2023.
9. Luo, Y., Xu, B., Cai, H., Bu, F., A Hybrid User Profile Model for Personalized Rec ommender System with Linked Open Data. *Enterprise Systems Conference (ES)*, Shanghai, pp. 243–248, 2014, doi: 10.1109/ES.2014.16.
10. Nahar, S., Pithawa, D., Bhardwaj, V., Rawat, R., Rawat, A., Pachlasiya, K., Quantum technology for military applications. in: *Quantum Computing in Cybersecurity*, pp. 313–334, 2023.
11. Shrivastava, M., Patil, R., Bhardwaj, V., Rawat, R., Telang, S., Rawat, A., Quantum Computing and Security Aspects of Attention-Based Visual Question Answering with Long Short-Term Memory. in: *Quantum Computing in Cybersecurity*, pp. 395–412, 2023.
12. Cremonesi, P., Koren, Y., Turrin, R., Performance of recommender algo rithms on top-n recommendation tasks, in: *Proc. of the fourth ACM conference on Recommender Systems*, ACM, New York, USA, pp. 39–46, 2010.
13. Bansal, N. and Koudas, N., Blogscope: A system for online analysis of high volume text streams, in: *Proceedings of the 33rd international conference on very large databases*, VLDB, pp. 1410–1413, 2007.
14. Fei Jiang, F.M., Pei, J., Chee Fu, A.W., Ix-cubes: iceberg cubes for data warehousing and olap on xml data, in: *CIKM '07: Proceedings of the sixteenth ACM conference on Information and knowledge management*, ACM, New York, NY, USA, pp. 905–908, 2007.
15. Chakravarthy, V.T., Gupta, H., Roy, P., Mohania, M., Efficiently linking text docu ments with relevant structured information, in: *VLDB '06: Proceedings of the 32nd International conference on very large data bases*, VLDB Endowment, pp. 667–678, 2006.

16. Lin, C.X., Ding, B., Han, J., Zhu, F., Zhao, B., Text Cube: Computing IR Measures for Multidimensional Text Database Analysis. *Eighth IEEE International Conference on Data Mining*, 2008.
17. Namdev, A., Patni, D., Dhaliwal, B. K., Parihar, S., Telang, S., Rawat, A., Potential Threats and Ethical Risks of Quantum Computing. in: *Quantum Computing in Cybersecurity*, pp. 335–352, 2023.
18. Shrivastava, M., Patil, R., Bhardwaj, V., Rawat, R., Telang, S., Rawat, A., Quantum Computing and Security Aspects of Attention-Based Visual Question Answering with Long Short-Term Memory, in: *Quantum Computing in Cybersecurity*, pp. 395–412, 2023.
19. Rawat, R., Chakrawarti, R.K., Raj, A.S.A., Mani, G., Chidambarathanu, K., Bhardwaj, R., Association rule learning for threat analysis using traffic analysis and packet filtering approach. *Int. J. Inf. Technol.*, *15*, 6, 3245–3255, 2023.
20. Sikarwar, R., Shakya, H. K., Kumar, A., Rawat, A., Advanced Security Solutions for Conversational AI. in: *Conversational Artificial Intelligence*, 287–301, 2024.
21. Patsariya, M. and Rajavat, A., A Progressive Design of MANET Security Protocol for Reliable and Secure Communication. *Int. J. Intell. Syst. Appl. Eng.*, *12*, 9s, 190–204, 2024.
22. Chauhan, D., Singh, C., Rawat, R., Chouhan, M., Conversational AI Applications in Ed-Tech Industry: An Analysis of Its Impact and Potential in Education, in: *Conversational Artificial Intelligence*, pp. 411–433, 2024.
23. Chauhan, D., Singh, C., Rawat, R., Dhawan, M., Evaluating the Performance of Conversational AI Tools: A Comparative Analysis, in: *Conversational Artificial Intelligence*, pp. 385–409, 2024.

8

An Investigation of Techniques to Encounter Security Issues Related to Mobile Applications

Devabalan Pounraj[1]*, Pankaj Goel[2], Meenakshi[3], Domenic T. Sanchez[4], Parashuram Shankar Vadar[5], Rafael D. Sanchez[6] and Malik Jawarneh[7,8]

[1]Department of Computer Science and Engineering, Bonam Venkata Chalamayya Engineering College (Autonomous), Dr. B. R. Ambedkar Konaseema District, Andhrapradesh, India
[2]Applied Science And Humanities, G. L. Bajaj Institute of Technology and Mangement, Gr. Noida, India
[3]Apeejay Stya University, Sohna, Haryana, India
[4]Cebu Technological University-NEC, City of Naga, Cebu, Philippines
[5]Yashwantrao Chavan School of Rural Development, Shivaji University, Kolhapur, Maharashtra, India
[6]Cebu Institute of Technology-University, Cebu, Philippines
[7]Oman College of Management and Technology, Muscat, Oman
[8]INTI International University, Subang, Malaysia

Abstract

Since its introduction, malware has been used to attack mobile devices. Fraudulent mobile applications and inserted harmful apps are the two main forms of standalone mobile malware assaults. A thorough grasp of the permissions stated in apps and API calls is essential if one wants to successfully defend against mobile malware's cyber risks. Permission requests and API calls are used in this study to create an effective categorization model. There are many APIs that Android applications utilize, thus to make it easier to detect malicious apps, the authors have come up with three alternative categorizing strategies: ambiguous, hazardous, and disruptive. Mobile malware often asks for harmful permissions to access sensitive data, as shown by the findings, which show that malicious apps use a different set of API calls than benign ones.

**Corresponding author*: devabalanme@gmail.com

Rajesh Kumar Chakrawarti, Ranjana Sikarwar, Sanjaya Kumar Sarangi, Samson Arun Raj Albert Raj, Shweta Gupta, Krishnan Sakthidasan Sankaran and Romil Rawat (eds.) Natural Language Processing for Software Engineering, (127–142) © 2025 Scrivener Publishing LLC

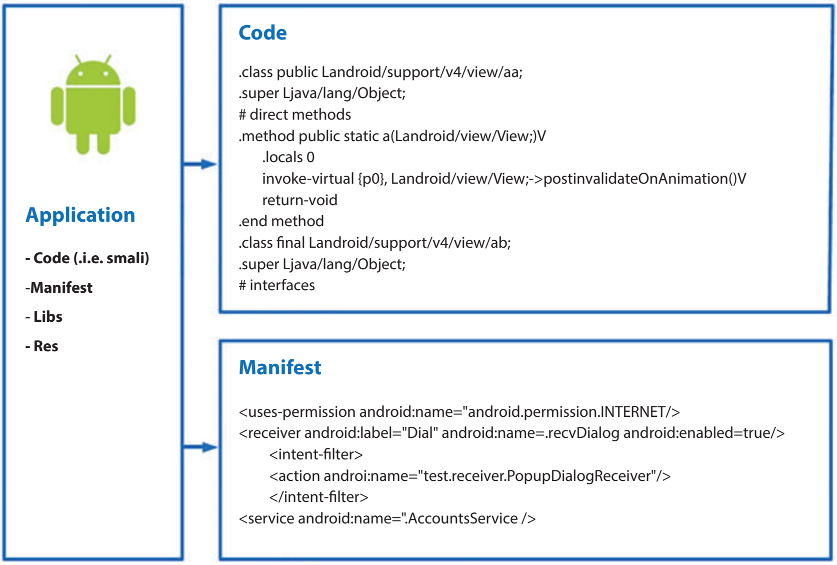
This article presents a comprehensive literature review of various methods to deal with Android malware and related security concerns. This article contains an analytical investigation of various methods used for tackling malwares in Android operating system. This study found that the machine learning algorithms like Support Vector Machine and Convolution Neural Network are most accurate in classification and prediction of malwares in Android operating system.

Keywords**:** Security, privacy, android, malware, counter measures, machine learning techniques, blockchain

8.1 Introduction

The Android malware detection challenge has been intensively explored by the academic security research community. On fixed datasets, machine learning approaches that have been developed in the past often have high detection rates. According to a few, the forecast times are also rather rapid. However, the criteria for malware detection go well beyond these figures of merit; therefore most of them are not adequate for real-world implementation [1]. Schematic representation is shown below in Figure 8.1.

As a result of the proliferation of Android apps and smartphones, the Internet of Things is rapidly becoming a reality. There has been a



```
.class public Landroid/support/v4/view/aa;
.super Ljava/lang/Object;
# direct methods
.method public static a(Landroid/view/View;)V
    .locals 0
    invoke-virtual {p0}, Landroid/view/View;->postinvalidateOnAnimation()V
    return-void
.end method
.class final Landroid/support/v4/view/ab;
.super Ljava/lang/Object;
# interfaces
```

```
<uses-permission android:name="android.permission.INTERNET/>
<receiver android:label="Dial" android:name=.recvDialog android:enabled=true/>
      <intent-filter>
      <action androi:name="test.receiver.PopupDialogReceiver"/>
      </intent-filter>
<service android:name=".AccountsService />
```

Figure 8.1 Schematic representation of android app.

considerable increase in the number of malicious software threats that are aimed at smartphones that run Android. In addition, owing to the widespread use of the Android operating system in Internet of Things devices, it is difficult to protect against attacks of this sort that are launched by malicious software. It is likely that the combination of machine learning methods and blockchain technology might improve the capabilities of Android Internet of Things devices to identify malware. Clustering, classification, and the blockchain are all components of the sequential process that was used in the introduction of the new technique. Clustering and classification techniques are used by machine learning to successfully capture malware data on the blockchain [2].

Malware [14] is always growing, both in terms of its diversity and its level of complexity, which makes it more difficult for detection technologies to keep up. The makers of malicious software use a broad variety of intricate methods whenever they wish to generate a new edition of the infection. To circumvent detection mechanisms and preserve the original functionality of the malicious code, these tactics include deleting, changing, or adding API calls that aren't really necessary [3].

Developers of malicious software [4] often focus their attention on the most widely used platforms in their efforts to infect the greatest number of devices feasible. The reason for this is that Android has been the subject of attacks on a consistently regular basis. The majority of efforts to battle malware on Android are focused on detection, with the objective of preventing infected devices from ever being infected in the very first place. For the purpose of automatically classifying malware samples and taking the appropriate measures to reduce the effect of those samples, this is a vital component.

It is possible to spread a significant number of malware samples on Android since they are variations of strains that are already known. Because of this, the security of an Android platform is dependent on a robust classifier that is able to manage large sample numbers. Automating variation detection and categorization is critical for detecting fraudulent activity and obtaining representative characteristics, particularly when dealing with larger datasets. Scalability plays a significant role in automating these processes, which is necessary for discovering fraudulent conduct. A deep neural network is made up of several layers, each of which has the capacity to learn and adjust its unique parameters to improve its classification accuracy. The computability of computers has been significantly improved by recent advancements in parallel computing and processing power [5, 6].

8.2 Literature Review

Malware identification is one of the several tedious and difficult parts of network security [1]. The proliferation of mobile devices and network approaches has greatly exacerbated the problem of malware, such as code obfuscation and zero-day attacks, on computer systems and networks. Methods for detecting malware that use machine learning (ML) have been suggested by several. The fast expansion of malware variants and assaults by adversarial examples (AEs) mean that research into malware detection is still continuing. Data visualization and adversarial training on detectors based on machine learning constitute our novel approach to the present constraint in malware detection. The ability to distinguish between various malware types will be greatly enhanced by this. With an average accuracy of 96.25% across all malwares evaluated, and a maximum accuracy of 97.33% according to experimental findings utilizing the MS BIG malware database and the Ember database, the suggested technique is able to thwart zero-day assaults. Malware assaults on the IT sector, particularly on AI and the Internet of Things, have increased noticeably. As the first general ML-based visualization tool for malware and variant detection, this work suggests Visual-AT, a universal binary format. To discover and evaluate variations of the same infection and previously undetected malware, another AT approach is used. This method makes use of visual data that has been altered by two ML algorithms. The experimental findings show that the proposed technique is more accurate and resilient than state-of-the-art ML-based malware detectors. Finally, the Visual-AT can accurately identify malware with a rate of up to 97.33% and an average detection rate of 96.56%. By reducing false positive rates by 81.17% and improving detection accuracy by 28.14%, Visual-AT outperforms both earlier and more modern machine learning-based visual identification methods.

It would be fascinating to see how the Visual-AT is used in novel situations, such as hierarchical labeling, given that the AEs are trained for every classification task using a horizontal label set (e.g. Animal-Dog-Poodle). Audio and voice processing are only two of the many possible applications of Visual-AT, which may make imposition and identification easier. Making the Visual-AT approach more relevant to other sorts of malware assaults, such the protection against the specialized attack termed Distributed Denial of Service (DDoS) [20–23], is another tough topic that requires study. Finally, there are a lot of ways the Visual-AT approach may boost machine learning [24–26] and computer security.

In this study, the authors address some of the most important issues that need to be considered before implementing Android malware detection systems in production environments [2]. A prospective approach must be contrasted with a constant stream of dynamic facts. A more accurate and realistic picture of detection performance may be obtained by simulating the continual flow of unknown file items into the classification process using streams of evolving data.

The authors created and used an ensemble method to detect Android malware automatically, taking into consideration the given real-world constraints. Because of its speed and reliability, Support Vector Machines use as inputs Atomic Naive Bayes classifiers that are built on several APK feature categories. Our example research successfully identified distinct forms of malware using different atomic classifiers, according to the authors. A model trained on 120,000 samples of a massive data stream has never previously had its results made publicly accessible.

The sensing, healthcare, and remote monitoring industries are just a few of the many that are benefiting from the Internet of Things (IoT) and its revolutionary changes to the environment [3]. Sending out a broadcast of all the malware data recorded in the blockchain's history makes it easy to spot any new infection. Producing weights for each set of characteristics, constructing an optimization parametric study, and continuously removing features with low weights that are redundant are all required steps in implementing the clustering strategy. The naive Bayes classifier is used to obtain the many properties of Android malware. The naive Bayes classifier use decision trees to glean additional crucial features for classification and regression, making it a very accurate and durable classifier. Our final strategy makes use of the permission blockchain to store real data acquired from extracted attributes in blocks of a distributed malware database, which speeds up and improves the accuracy of run-time malware detection.

Malware that targets Android phones has also proliferated in tandem with the exponential growth of Android smartphones [4]. One possible aspect is that third-party programs made for Android may run on any Android device, regardless of the OS. A key differentiator of Android is its inter-process communication capability, which enables components to be reused between processes. This is a great approach for the Android framework to access a lot of third-party services. Commonly, this kind of communication is managed by binding messaging objects called intents during runtime. Permissions and other well-studied characteristics may not be able to capture viral intentions as precisely as intents, which provide semantically rich information. This decision will be the worst one to make in the long run. You may maximize its potential by combining it with other well-known features. Out of a total of 7,406 applications, 1846 were found to be clean and 5,5600 to be contaminated. Among Android permissions, 83% could be identified and 91% of Android Intents could be identified. Using both features together raises the detection probability to 95.5%.

To deduce malicious activity, this study suggests novel algorithms that use associative principles and are based on recurrent subsequence alignment. The proposed technique [5] considers the chronology and probability of transitioning between two API invocations in the call sequence to detect groups of malwares that have similar dangerous behavior. It is not required that these subsets be sequential. Because they may be used in dynamic analysis scenarios, the following malware classification method is resistant to obfuscation and evasion strategies based on API call perturbation. The results show that it beats two well-known methods for malware detection: one that uses Markov chains and another that makes use of API call sequence matching techniques. In this experimental assessment, the suggested strategy outperformed its rivals in terms of categorization performance.

In this paper, the authors showcase the CANDYMAN approach to Android malware detection, which combines dynamic analysis with Markov chains. Using Markov chains and dynamic analysis approaches, it is possible to extract useful information from malware samples in the form of an ordered series of states. With this information, we may design features that can be used to classify things. By using Deep Learning algorithms and unbalanced learning approaches, among other machine learning techniques, the authors train a library of malware samples from various families. After that, they assess the proposed method. The Drebin dataset originally had 179 malware families; however, 4,442 samples were selected from 24 distinct families due to a lack of appropriate and relevant samples. The validity and relevance of these samples were further assessed. The experimental results in this dataset demonstrated an exceptionally high rate of accuracy, measuring 81.8% [6].

There are a lot of Android applications that track trends, customer preferences, and industry news [7]. The open-source nature of Android makes it possible for unlicensed app shops, and the platform's lack of app screening makes it inevitable that harmful software will infect Android-based smartphones and tablets. In response, specialists have devised a strategy that safeguards Android users against harmful applications by constructing a model that use ensemble Rotation Forest (RF). They utilized a dataset of 2,130 samples to test the plan's efficacy. An accuracy rate of 88.56% and a sensitivity of 88.10% we both attainable with the proposed approach.

The suggested method's accuracy was tested using a widely-used support vector machine (SVM) model. Its accuracy was found to be 3.33% higher than that of the SVM model. The outcomes of the tests demonstrated that this approach shows great promise for detecting malware on Android devices.

Users of the Android operating system have been somewhat worried about data security since its debut. Whether or not an app's special features can withstand viral attacks is highly dependent on those features. If Android malware detection is an issue, a Siamese network might be a solution. Feature grouping is an alternative to traditional methods for choosing the most appropriate API requests and permissions. We provide a Feature Centralized Siamese Convolutional Neural Network (FCSCNN) to speed up the identification process. A database's mean centers for benign and malicious data are first retrieved using the FCSCNN. We can find out what category an application belongs in by comparing its distance from the two centers. Findings from the "VirusShare" dataset show that the suggested method, FCSCNN, can detect viruses in less than 0.1 second with an accuracy of 98.07 percent [8]. An F-measure of 94.3% was achieved when the suggested strategy was used to a real-world dataset of 27,891 Android applications [9].

A technique for identifying Android malware using features collected from instruction call graphs is proposed in this work [10] based on previous research. When presented with a balanced dataset, deep neural networks make greater use of their exploration capabilities to distinguish between safe and hazardous execution routes. There isn't a publicly accessible model for Android malware detection available right now. A synthetic data set is used in this study. To find the best values for statistical indicators, we use a grid search strategy that takes into account both network parameters and hyper-parameters. We put the proposed technique through its paces using a dataset consisting of 24,650 negative samples and 25,000 positive ones. When comparing baseline classifiers with metrics for deep network architecture statistics, runtime is one of the metrics employed. A table displays the outcomes. A total of 91.42 percent accuracy, with an extra 91.11 percent F-measured, is achieved by the malware detection service that is supplied, according to the research.

Being able to detect malware is crucial for smart device security [11]. Despite its inability to identify zero-day assaults or polymorphic infections, signature-based approaches continue to be widely used. A hybrid approach is recommended for Android malware detection. This post will provide a strategy for analyzing Android apps that integrates heuristic and signature-based approaches. In this context, "reverse engineering" refers to disassembling an Android app to get its manifest and binary files.

The authors extracted manifest and binary files, as well as identified Android malware, using state-of-the-art machine learning methods. Using a battery of tests, they put a suite of classifiers through their paces to get there. When it comes to recognizing binary files, SVM is the way to go, but when it comes to identifying manifest.xml files, KNN is where it's at. The methodology demonstrated promising outcomes in malware identification when tested on benchmark datasets.

The authors of this paper [12] provide a new method for detecting malware that makes use of convolution recurrent neural networks and opcode sequences. Executable files are statistically defined as a collection of machine codes. Malware detection using opcode sequences is a theoretical possibility. A malware detection system that uses deep learning using these sequences as input and a method for extracting opcode sequences have been described in detail. The next stage involves a front-end convolutional auto encoder operating at the opcode level and a back-end dynamic neural network classifier. The experimental results showed that the suggested model had a 95% TPR, a receiver operating characteristic-area under the curve of 0, and a detection accuracy of 96%. Opcode sequences were the gold standard for malware detection back then, with a 97% success rate and an 82% false positive rate. The suggested model's accuracy was 96%, which was somewhat lower than the current method, but its TPR was 95%, which was much higher. The suggested approach is more effective at detecting harmful software because of this.

Customers are at risk from zero-day malware [13] samples because no safeguards are in place to prevent freshly found, unexpected activity. To deal with zero-day malware, prior information is essential for malware detection. The experts agree that it's bad when a once-dangerous feature becomes its real form. This implies that when confronted with new dangers, tightly linked feature-engineering that uses previous domain expertise may not work. In lieu of human specialists, this study employs a deep learning neural network that is completely ignorant of any possibly harmful characteristics. Examples of these traits include app permissions and the usage of hidden Android APIs. Three significant advantages are associated with our method. A malware detector for Android has been created using AI and machine learning. It can identify, rank, and categorize threats without the need for malware domain knowledge. The authors used Drebin and AMD benchmarks to assess our model's functionality in zero-day scenarios. Compared to state-of-the-art approaches, the proposed model achieved detection rates as high as 81% and 91%, respectively, a 57% improvement. With F1 scores of 0.9928 on the Drebin dataset and 0.9963 on the AMD benchmark dataset, respectively, overall detection performance was improved by an average of 77% when compared to current best practices. The findings were further contextualized by this.

An open vulnerability exists in the Android operating system because malicious programs (also termed malware) often find methods to circumvent the security mechanisms, and consumers are unaware of malware before downloading an app. Many techniques based on machine learning have been developed to identify malware in data collected via static analysis in an effort to address this issue. The authors summarized previous research on the topic and test the efficacy of supervised machine learning methods using the Drebin dataset's static analysis data. With the use of six well-known classification approaches, each with its own unique set of characteristics, they want to identify Android malware and features. With a small set of features, the proposed model was able to obtain good classification accuracy.

A meteoric rise in malware assaults has accompanied the widespread use of cellphones. When viruses use covert strategies, signature-based approaches struggle to detect them [15]. This study introduces the PIndroid framework, a tool for detecting malicious Android applications. When it comes to anti-malware solutions, PIndroid is the only one that uses Ensemble approaches in combination with permissions and intents. The suggested solution had the highest reported success rate of 99.8 percent across 1,745 real-world applications. The suggested method seems to be successful in detecting harmful applications, according to the empirical data.

Hackers have often targeted Android due to its widespread use and ease of access. Hackers are constantly developing new methods of attack and finding new security flaws in Android as Google rolls out new versions of the operating system. The security and resource restrictions of Android provide unique obstacles for malware attack detection and investigation. The authors of this work suggest several novel behavioral methods for classifying and detecting Android malware [16]. Decompiling the Android malware dataset allows one to construct an encoded list of potentially malicious API classes and functions. Following the generation of several sequence alignments for various malware types utilizing the encoded patterns, the profile hidden Markov model is used. An unknown program's potential danger may be assessed using a log probability score. Outperforming state-of-the-art approaches to Android malware detection, this framework achieves a remarkable 94.5% detection accuracy.

There has been a lot of buzz lately around Android malware detection [17]. Models for mobile virus detection are often constructed using machine learning approaches. Both static and dynamic application information may be gathered by these models. Hopefully, a far broader range of static and dynamic properties will be retrievable. Malware data is famously hard to collect owing to its intrinsic variety and its high dimensionality. It is possible to identify zero-day malware with the use of unsupervised malware detection technologies. In this scenario, an unsupervised feature reduction approach might be used to lower the data's dimensionality. This paper proposes a feature learning approach called Subspace based Restricted Boltzmann Machines (SRBM) to potentially reduce the dimensionality of virus detection data. Initially, the initial data set is searched for several subspaces. In addition, every subspace is given its own unique RBM. By combining the results of the hidden layers of the trained RBMs, a lower-dimensional data representation is produced. The features learnt by SRBM outperform those by other feature reduction methods when performance is assessed using clustering evaluation metrics. Datasets from OmniDroid, CIC2019, and CIC2020 show that NMI, ACC, and Fscore are effective.

Four malware detection methods—k-medoid-based next-neighbors (KMNN), all nearest neighbors (ANN), weighted all nearest neighbors (WANN), and first-closest (FNN)—use Hamming distance to identify sample similarity [18]. This approach has the potential to stop the spread of malware via an alarm system. There is a wealth of information here on detection techniques and the algorithms that go along with them. To verify their usefulness, a thorough evaluation of suggested similarity-based detection methods is carried out. Our analysis makes use of three datasets, including Drebin and malware as well as Contagio and Genome, which include both high-quality and low-quality Android applications. Classifier's efficacy was shown by comparing its results with those of the FalDroid methods, the PDME (program dissimilarity measure based on entropy), and a handful of other state-of-the-art algorithms. Several factors, such as API, purpose, and authorization, were examined on these three datasets. The data shows that the suggested algorithms are state-of-the-art with an accuracy rate of over 90% and, under some circumstances (such as when taking API features into account), over 99.99 percent.

The Android operating system has caused a tsunami wave of change in the mobile and portable device industries. This free and open-source mobile platform runs on virtual computers and can support a diverse ecosystem with millions of devices. There has been a meteoric rise in the amount of malware aimed at Android devices; at present, 99 percent of all malware detected on smartphones is Android-specific. There are other methods for evaluating and detecting malicious Android apps that have been published in academic journals. Because Android malware has evolved and proliferated so rapidly in recent years, it has become more difficult to identify new types of Android malware. Our 'End to End Deep Learning Architectures that detect and attribute Android malware using opcodes retrieved from application bytes' [19] are specialized learning models that have been designed to handle this challenge. By using neural networks with bidirectional long short-term memory (BiLSTMs), which outperform the current state-of-the-art, it is possible to identify static behavior in Android malware. Recurrent neural networks, Long Short-Term Memory networks, and their Bidirectional variant are the neural networks used in our study. For static malware analysis, our study also makes use of conventional neural architectures like Diabolo networks (autoencoders), deep convnets, and generative graphical models like deep belief networks. There is now a better bytecode dataset thanks to the inclusion of three publicly available and autonomously updated bytecode datasets. Results showed an F1-score of 0.996 and an accuracy of 0.999 on a dataset of over 1.8 million Android applications.

To keep up with the rapid growth of the IoT, cyber-physical systems (CPS) that use IoT principles to provide a range of services have become a focus for malware research and detection. Malware detection and analysis have already achieved great strides with the use of cutting-edge machine learning techniques like deep learning. But there are some problems with this method as well. The authors claim that certain methods aren't as reliable as they should be due to the presence of noise and outliers in the available malware datasets. The conclusion is that malware categorization might need some work. To get around this issue, it could be conceivable to combine correntropy with deep learning models. Below, we outline our technique [20] for malware detection and analysis, which involves reconstructing the loss function of a popular deep learning model called the Convolutional Neural Network (CNN). This allows us to find outlier behaviors. For challenging datasets with a high noise level, the mixed correntropy is a useful similarity metric to use. Results from testing the suggested method on both a popular benchmark dataset and a real-world malware dataset demonstrate its efficacy in learning.

8.3 Results and Discussions

Malware detection is one of the most difficult parts of network security. Anti-malware detection approaches that are based on machine learning (ML) have been suggested. The suggested strategy is able to avoid the zero-day attack and achieve up to 97.73 percent accuracy.

There has been an increase in the amount of malware targeting Android phones. Intents, which are late runtime binding messaging objects, are often responsible for this communication. An overall number of 7406 applications were examined, of which 1846 were found to be clean and 5560 contaminated. A total of 91% of Android Intents were identified, while an equal amount of 83% of Android permissions were identified. It can be identified 95.5% of the time when both traits are utilized in conjunction with one another.

To evaluate the suggested approach, machine learning techniques such as imbalanced learning methods and Deep Learning algorithms are utilized to train a dataset of malware samples from a range of families. According to the Drebin dataset, 179 separate malware families were represented, resulting in 4,442 samples divided among 24 distinct malware families, which were then chosen for relevance and validity owing to a scarcity of relevant and valid samples. The experimental findings reveal an accuracy rate of 81.8 percent for this dataset, which is an extremely high rate for this dataset [6].

The FCSCNN starts by determining the benign and malicious mean centers in the database. The distance between an application and the two centers is measured to determine which class to give it. The "VirusShare" dataset has been used to demonstrate that the proposed strategy FCSCNN is 98.07 percent accurate and takes less than 0.1 second to identify infections [8].

Our proposed approach for detecting mobile malware was shown to be successful using a real-world dataset of 27,891 Android apps, with an F-measure of 94.3%. Our approach to malware forensics and mobile app analysis, we feel, will be very valuable [9].

8.4 Conclusion

Malware has been used to target mobile devices since they were first introduced. The two most common types of independent mobile malware attacks are fraudulent mobile applications and malicious apps that are embedded within legitimate mobile applications. If one wishes to effectively protect against mobile malware's cyber threats, one must have a comprehensive understanding of the permissions provided in applications and API requests. Using permission requests and API calls, this research aims to develop an efficient classification model. Because there are several APIs that Android applications make use of, the authors have devised three additional categorization methodologies to make it simpler to identify malicious applications: confusing, dangerous, and disruptive, among others. According to the research, mobile malware often requests detrimental rights to access sensitive data, as shown by the fact that malicious applications make use of a distinct set of API calls than benign apps. An in-depth examination of the literature on several ways for dealing with Android malware and associated security problems is presented in this article. This article is a comprehensive look at how other people have dealt with Android malware and other security issues. This article is about different ways to deal with malware in the Android operating system. This study found that machine learning algorithms like the Support Vector Machine and the Convolution Neural Network are the best at classifying and predicting malware in the Android OS.

References

1. Liu, X., Lin, Y., Li, H., Zhang, J., A novel method for malware detection on ML-based visualization technique. *Comput. Secur.*, 89, 101682, 2020, Available: 10.1016/j.cose.2019.101682 [Accessed 6 January 2022].
2. Palumbo, P., Sayfullina, L., Komashinskiy, D., Eirola, E., Karhunen, J., A pragmatic android malware detection procedure. *Comput. Secur.*, 70, 689–701, 2017, Available: 10.1016/j.cose.2017.07.013 [Accessed 6 January 2022].
3. Kumar, R., Zhang, X., Wang, W., Khan, R., Kumar, J., Sharif, A., A Multimodal Malware Detection Technique for Android IoT Devices Using Various Features. *IEEE Access*, 7, 64411–64430, 2019, Available: 10.1109/access.2019.2916886.
4. Feizollah, A., Anuar, N., Salleh, R., Suarez-Tangil, G., Furnell, S., AndroDialysis: Analysis of Android Intent Effectiveness in Malware Detection. *Comput. Secur.*, 65, 121–134, 2017, Available: 10.1016/j.cose.2016.11.007 [Accessed 6 January 2022].
5. D'Angelo, G., Ficco, M., Palmieri, F., Association rule-based malware classification using common subsequences of API calls. *Appl. Soft Comput.*, 105, 107234, 2021, Available: 10.1016/j.asoc.2021.107234 [Accessed 6 January 2022].
6. Martín, A., Rodríguez-Fernández, V., Camacho, D., CANDYMAN: Classifying Android malware families by modelling dynamic traces with Markov chains. *Eng. Appl. Artif. Intell.*, 74, 121–133, 2018, Available: 10.1016/j.engappai.2018.06.006 [Accessed 6 January 2022].
7. Zhu, H., You, Z., Zhu, Z., Shi, W., Chen, X., Cheng, L., DroidDet: Effective and robust detection of android malware using static analysis along with rotation forest model. *Neurocomputing*, 272, 638–646, 2018, Available: 10.1016/j.neucom.2017.07.030 [Accessed 6 January 2022].

8. Kong, K., Zhang, Z., Yang, Z., Zhang, Z., FCSCNN: Feature centralized Siamese CNN-based android malware identification. *Comput. Secur.*, 112, 102514, 2022, Available: 10.1016/j.cose.2021.102514 [Accessed 6 January 2022].
9. Alazab, M., Alazab, M., Shalaginov, A., Mesleh, A., Awajan, A., Intelligent mobile malware detection using permission requests and API calls. *Future Gener. Comput. Syst.*, 107, 509–521, 2020, Available: 10.1016/j.future.2020.02.002 [Accessed 6 January 2022].
10. Pektaş, A. and Acarman, T., Learning to detect Android malware via opcode sequences. *Neurocomputing*, 396, 599–608, 2020, Available: 10.1016/j.neucom.2018.09.102 [Accessed 6 January 2022].
11. Rehman, Z. *et al.*, Machine learning-assisted signature and heuristic-based detection of malwares in Android devices. *Comput. Electr. Eng.*, 69, 828–841, 2018, Available: 10.1016/j.compeleceng.2017.11.028 [Accessed 6 January 2022].
12. Jeon, S. and Moon, J., Malware-Detection Method with a Convolutional Recurrent Neural Network Using Opcode Sequences. *Inf. Sci.*, 535, 1–15, 2020, Available: 10.1016/j.ins.2020.05.026 [Accessed 6 January 2022].
13. Millar, S., McLaughlin, N., Martinez del Rincon, J., Miller, P., Multi-view deep learning for zero-day Android malware detection. *J. Inf. Secur. Appl.*, 58, 102718, 2021, Available: 10.1016/j.jisa.2020.102718 [Accessed 6 January 2022].
14. Syrris, V. and Geneiatakis, D., On machine learning effectiveness for malware detection in Android OS using static analysis data. *J. Inf. Secur. Appl.*, 59, 102794, 2021, Available: 10.1016/j.jisa.2021.102794 [Accessed 6 January 2022].
15. Idrees, F., Rajarajan, M., Conti, M., Chen, T., Rahulamathavan, Y., PIndroid: A novel Android malware detection system using ensemble learning methods. *Comput. Secur.*, 68, 36–46, 2017, Available: 10.1016/j.cose.2017.03.011 [Accessed 6 January 2022].
16. Sasidharan, S. and Thomas, C., ProDroid — An Android malware detection framework based on profile hidden Markov model. *Pervasive Mob. Comput.*, 72, 101336, 2021, Available: 10.1016/j.pmcj.2021.101336 [Accessed 6 January 2022].
17. Liu, Z., Wang, R., Japkowicz, N., Tang, D., Zhang, W., Zhao, J., Research on unsupervised feature learning for Android malware detection based on Restricted Boltzmann Machines. *Future Gener. Comput. Syst.*, 120, 91–108, 2021, Available: 10.1016/j.future.2021.02.015 [Accessed 6 January 2022].
18. Taheri, R., Ghahramani, M., Javidan, R., Shojafar, M., Pooranian, Z., Conti, M., Similarity-based Android malware detection using Hamming distance of static binary features. *Future Gener. Comput. Syst.*, 105, 230–247, 2020, Available: 10.1016/j.future.2019.11.034 [Accessed 6 January 2022].
19. Amin, M., Tanveer, T., Tehseen, M., Khan, M., Khan, F., Anwar, S., Static malware detection and attribution in android byte-code through an end-to-end

deep system. *Future Gener. Comput. Syst.*, 102, 112–126, 2020, Available: 10.1016/j.future.2019.07.070 [Accessed 6 January 2022].

20. Luo, X., Li, J., Wang, W., Gao, Y., Zhao, W., Towards improving detection performance for malware with a correntropy-based deep learning method. *Digit. Commun. Netw.*, 7, 4, 570–579, 2021, Available: 10.1016/j.dcan.2021.02.003 [Accessed 6 January 2022].
21. Nahar, S., Pithawa, D., Bhardwaj, V., Rawat, R., Rawat, A., Pachlasiya, K., Quantum Technology for Military Applications, in: *Quantum Computing in Cybersecurity*, pp. 313–334, 2023.
22. Pithawa, D., Nahar, S., Bhardwaj, V., Rawat, R., Dronawat, R., Rawat, A., Quantum Computing Technological Design Along with Its Dark Side, in: *Quantum Computing in Cybersecurity*, pp. 295–312, 2023.
23. Shrivastava, M., Patil, R., Bhardwaj, V., Rawat, R., Telang, S., Rawat, A., Quantum Computing and Security Aspects of Attention-Based Visual Question Answering with Long Short-Term Memory, in: *Quantum Computing in Cybersecurity*, pp. 395–412, 2023.
24. Rawat, R., Chakrawarti, R.K., Sarangi, S.K., Patel, J., Bhardwaj, V., Rawat, A. and Rawat, H. eds. *Quantum Computing in Cybersecurity*. John Wiley & Sons, 2023. https://onlinelibrary.wiley.com/doi/book/10.1002/9781394167401
25. Patsariya, M. and Rajavat, A., A Progressive Design of MANET Security Protocol for Reliable and Secure Communication. *Int. J. Intell. Syst. Appl. Eng.*, *12*, 9s, 190–204, 2024.
26. Rathi, M. and Rajavat, A., Investigations and Design of Privacy-Preserving Data Mining Technique for Secure Data Publishing. *Int. J. Intell. Syst. Appl. Eng.*, *11*, 9s, 351–367, 2023.

9

Machine Learning for Sentiment Analysis Using Social Media Scrapped Data

Galiveeti Poornima[1]*, Meenakshi[2], Malik Jawarneh[3,4], A. Shobana[5], K.P. Yuvaraj[6], Urmila R. Pol[7] and Tejashree Tejpal Moharekar[8]

[1]*Presidency University, Bangalore, India*
[2]*Apeejay Stya University Sohna Haryana, Haryana, India*
[3]*Oman College of Management and Technology, Muscat, Oman*
[4]*INTI International University, Subang, Malaysia*
[5]*Department of Computer Applications, Sri Krishna Arts and Science College, Coimbatore, India*
[6]*Department of Mechanical Engineering, Sri Krishna College of Engineering and Technology, Coimbatore, India*
[7]*Department of Computer Science, Shivaji University, Kolhapur, India*
[8]*Yashwantrao Chavan School of Rural Development, Shivaji University, Kolhapur, Maharashtra, India*

Abstract

Social media encompasses a wide range of internet apps that enable users to share, find, and engage with material created by other users. The number of persons utilizing social media increases by up to 3.80 million on a daily basis. By the start of 2020, the number of internet users is projected to reach 1 billion. Social media is a significant catalyst for acquiring and disseminating information in several domains such as entertainment, commerce, science, politics, and crisis management. It enables users to publish and share a diverse range of media formats, including text, videos, pictures, and audio. By conducting data analysis on social media, a person can access a wide range of information, including trends, concerns, and key individuals. Sentiment Analysis (SA) aims to ascertain the emotional response of individuals towards a specific service, business, or product. Sentiment analysis utilizes a diverse range of approaches, strategies, and tools to identify and extract subjective information, such as views and attitudes, from a given language.

**Corresponding author*: varaprabha.lita@gmail.com

Rajesh Kumar Chakrawarti, Ranjana Sikarwar, Sanjaya Kumar Sarangi, Samson Arun Raj Albert Raj, Shweta Gupta, Krishnan Sakthidasan Sankaran and Romil Rawat (eds.) Natural Language Processing for Software Engineering, (143–154) © 2025 Scrivener Publishing LLC

Traditionally, SA has sought to assess any information that is accessible to the public on the Internet. When utilized in conjunction with theme analysis, sentiment analysis has the capability to accurately detect both positive and negative emotions. Deep Learning Techniques (DLTs) are valuable in Sentiment Analysis (SA) because they have the ability to teach both supervised and unsupervised categories. DLTs comprise several network types such as feature presentation, phrase modeling, text creation, word representation estimation, vector representation, and sentence classification.

***Keywords*:** Sentiment analysis, deep learning social media scrapped data, twitter data, accuracy, CNN, SVM, AdaBoost

9.1 Introduction

The most vivid representation of a person's deeds is found in their opinions. Subjective feelings such as viewpoint, sentiment, attitude, and so on had a significant impact on the actions of humans. The decisions and convictions that we make are influenced by the way in which other people view and feel about the world. When it comes to making decisions, we frequently consult with one another to get their perspectives on the matter. Considering that the opinion of the general public regarding their goods and services is always significant to them, this is something that can be beneficial to both individuals and a big number of organizations [1].

Even the typical consumer will examine reviews published by individuals who have already purchased a product before making a final choice about whether or not to purchase it. Customers' reviews and star ratings are shown next to each product on online marketplaces such as Amazon.com. These reviews are helpful in providing the buyer with recommendations of things to purchase. The utilization of social media platforms, such as Facebook, Twitter, and blogs, among others, for the purpose of conveying desires, accusations, and feelings regarding any service, product, or topic has witnessed a significant increase among users of a significant number of different countries. It is possible for a person to convey his/her opinion in connection to many different things, including an event, a person, a product, or even an organization. Components and features are examples of possible parts that make up an entity. A phone, for example, is an example of a thing. There are several components that make up the mobile device known as a "phone," including the battery, the screen, and so on. The quality of the camera and the voice are two examples of the traits that it possesses. Either by directly addressing any entity or by drawing parallels between them, individuals have the ability to communicate their thoughts in a number of different ways. Take into consideration the following example to better demonstrate the distinction between the two as follows: "battery life of Nokia is more than iPhone." In the first scenario, the speaker is expressing his/her viewpoint regarding the "battery life" feature of a Nokia phone. In the second scenario, he/she are comparing the two companies, Nokia and iPhone, and the battery life features that each of them possesses [2].

Despite the fact that surveys, interviews, and other antiquated techniques of acquiring and assessing opinionated data are still utilized, sentiment analysis by hand is an unattainable undertaking due to the vast volume of opinionated data and the absence of structure in this data. Methods that are capable of automatically extracting sentiment from unstructured opinionated data are required as a result of this. To sort through this mountain of subjective data, we are employing a wide range of Natural Language Processing (NLP) activities. Particularly, sentiment analysis is gaining popularity due to its objective of polarity classification. This objective can be accomplished through binary or multiclass classification, and it generates outcomes such as positive or negative, neutral or thumbs up or thumbs down. The application of sentiment analysis (SA) to the study of public opinion and sentiment exposes the influence on individuals as well as the impact on the community as a whole during the course of the research. Opinions are the primary foci of sentiment analysis [3].

As a result of the flood of biased data, sentiment analysis has developed into a significant area of research in recent years. Simply put, it is extremely important for the advancement of artificial intelligence. In particular, the meteoric expansion of recommendation websites, blogs, and social networking platforms such as Twitter and Facebook has attracted the interest of a great number of organizations in the field of sentiment analysis. When it comes to finding new customers and successfully selling their products, businesses and corporations rely on this kind of subjective information [4]. A wide range of user-generated content (UGC) that may be found on the internet, including blog posts, reviews, comments, and other forms of online material, can be subjected to automated sentiment analysis. To put it another way, it analyzes a piece of text to determine how people are feeling and then assigns a label to that feeling, such as positive, negative, neutral, furious, sad, etc. To summarize, Sentiment Analysis, which is sometimes referred to as Opinion Mining, is a method that is utilized in the field of machine learning (ML) that employs text analysis, computational linguistics, and natural language processing to classify the feelings that individuals have regarding any certain matter [5].

9.2 Twitter Sentiment Analysis

Analysis of sentiment has been the primary focus of studies for a considerable amount of time. Students in academic institutions have shown an interest in sentiment analysis as a method for determining the viewpoints of the general public on a number of different subjects. It has become an indispensable instrument, whether it be for a single individual or for a major organization that is making a decision based on opinionated data that is available to the public. A rising number of companies, including the government, are incorporating sentiment analysis into their marketing strategies and are offering financial assistance for projects that are relevant to this field. The various types of sentiment research that may be performed on social media can also lead to wonderful business opportunities. Systems that provide recommendations and models that are based on customer relationships stand to gain a great deal from its deployment. For the purpose of illustrating the argument, it is helpful for recommendation systems to determine which features customers appreciate and for them to eliminate features that elicit negative feedback. Despite the fact that there has been a consistent interest in sentiment analysis for a considerable amount of time, the field has only recently witnessed a significant increase in the amount of study [6].

Since its inception in the 1950s, sentiment analysis has been used for the goal of extracting opinions from written texts. On the other hand, the explosion of web 2.0 itself is directly responsible for the extensive use of sentiment analysis to opinionated data found online. When compared to the extensive histories of linguistics and Natural Language Processing (NLP), the amount of study conducted on opinions and sentiments before to the year 2000 was rather unrepresentative. Prior to the widespread recognition of sentiment analysis, the majority of the older publications focused their attention on examining a wide range of themes, including affects, metaphors, points of view, and a variety of other subjects. Prior to the introduction of the World Wide Web (WWW), there may have been a scarcity of data that was available digitally and contained expressions of opinion. People have turned to the time-honored method of word-of-mouth (WOM) to learn about people's ideas and feelings regarding any given topic or situation. This is due to the fact that the internet has not advanced technologically [7].

Word-of-mouth marketing, often known as WOM, was defined by S. Tokes and Lomax as "interpersonal communication regarding product or services regardless of whether the receiver regards communicator as impartial." Before the World Wide Web became widely known, organizations would traditionally rely on polls, surveys, and interviews to gather information about people's opinions. The subject of study known as sentiment analysis has always been one that is both fascinating and dynamic. However, with the advent of web 2.0, individuals have never before had access to a platform that is unparalleled in its ability to express their insightful opinions on a broad variety of topics. Numerous factors are contributing to the rising interest that researchers are displaying in the field of sentiment analysis. It is possible that one of the contributing factors is the fact that sentiment analysis offers a substantial number of potential applications across a wide range of industries, such as healthcare, politics, social events, finance, and many others. By making good use of the vast volumes of sentiment-bearing opinionated data that are available to the public, organizations and government enterprises have the potential to significantly benefit from accessing this information. Another, more obvious reason is that this is the case. It has been observed that the rise of social media technology has coincided with the emergence of sentiment analysis as a subject of study [8].

When it comes to the dissemination of information and the expression of opinions, Twitter stands out as the social media platform with the most users. Through the use of this casual platform, individuals from all over the world are able to send and receive messages. It is because of this that a lot of natural disasters have been brought to the notice of the general population. It all began in 2006 as a platform for sending and receiving text messages (SMS), but since then, it has expanded significantly beyond its intended purpose. This is largely due to the efforts of social media influencers and well-known celebrities who use it to engage with their individual audiences, whether for personal or business purposes. According to research (https://learn.g2.com/Twitter-statistics), there are more than 321 million individuals who use Twitter on a daily basis, and they transmit an average of 500 million tweets with their accounts. This day and age, Twitter remains one of the most popular microblogging services due to the fact that millions of users submit millions of tweets every single day. There is an opportunity for brand expansion on Twitter due to the fact that businesses are able to communicate with and gain a better understanding of their customers through the platform's opinionated data. The fact that more than 80 percent of the material is not organized is a further insult. "Tweet" is the term used to describe a message that is posted by an individual on the microblogging service Twitter. Given that, the maximum number of characters that can be included in a tweet is 160. As a consequence of this, tweets are written in a manner that is less formal [9, 10].

There is a significant amount of interest in the study of Twitter sentiment analysis at the present time. This information can be utilized by both the government and corporations to evaluate the effectiveness of their systems in the midst of the flooding of subjective data that is freely available on the internet [18–21]. Twitter is considered to be one of the most popular microblogging websites, where users frequently communicate their thoughts, opinions, and attitudes regarding a wide range of topics and events. Twitter is the most popular social networking platform [22–24], and it is used by not only the regular person but also politicians and celebrities. The elections for the Lok Sabha, the decision regarding the Ram Mandir, the decision regarding the shutdown, and the decision regarding demonetization are just a few instances of the events that cause Twitter to become flooded with tweets that are related to the current topic. It is enlightening to observe how individuals are feeling about a variety of subjects and current events based on the tweets that they have released [11].

In spite of the fact that there is a plethora of opinionated content available on Twitter in the form of tweets, manually gathering the vast amount of data, extracting relevant information, and then synthesizing the opinions that have been gathered into a meaningful manner is a big job. What this demonstrates is the importance of using software that analyzes sentiment. Despite the fact that opinionated data in the form of tweets is relatively unstructured, group thoughts from a variety of persons are more illuminating than individual perspectives. In addition, the vast majority of the digital data that can be accessed online is unstructured, meaning that it is not organized in any specific way (https://learn.g2.com/structured-vs-unstructured-data). The distillation and extraction of relevant information from data that carries sentiment [25] is required to improve the quality of decisions that are made for society as a whole. Sentiment analysis on Twitter can be of use in this regard.

According to [12], the basic objective of Twitter sentiment analysis is to extract the feelings that are associated with a text, which in this case is tweets.

Analysis of Twitter sentiment has been the subject of a great number of previous research, which have utilized a wide range of features and approaches (including lexicon-based, machine learning, and hybrid ways of thinking). The work that researchers have done on Twitter sentiment analysis has achieved considerable results, according to their reports.

To further extract the essential information from opinionated data, a number of methods for sentiment analysis [13] have been developed. The vast majority of research projects that have been conducted on Twitter sentiment analysis have utilized supervised machine learning approaches.

These techniques entail training classifiers with data that has been labeled. Despite this, there is a dearth of research that compares a large number of various state-of-the-art classifiers that make use of different characteristics. Through a comparison of the performance of a number of state-of-the-art classifiers in respect to features, the major objective of this thesis is to offer a solution to the problem that has been presented. The objective is to determine which classifiers offer the greatest degree of success when applied to particular sets of characteristics. [14–19] Several classifiers will be evaluated on several domains, including a public benchmark dataset and a real-time Twitter dataset (in which real-time tweets are obtained from Twitter). These datasets will be used to test the algorithms.

9.3 Sentiment Analysis Using Machine Learning Techniques

The methods that were used may be seen in Figure 9.1.

This section consists of three classification techniques: Support Vector Machine, AdaBoost, and Convolutional Neural Network.

Support Vector Machine (SVM) is an excellent choice to study if you are seeking for an algorithm that may assist you in recognizing pedestrians. A common method for displaying images that are produced by computer vision algorithms is to use a non-linear matrix that is composed of square

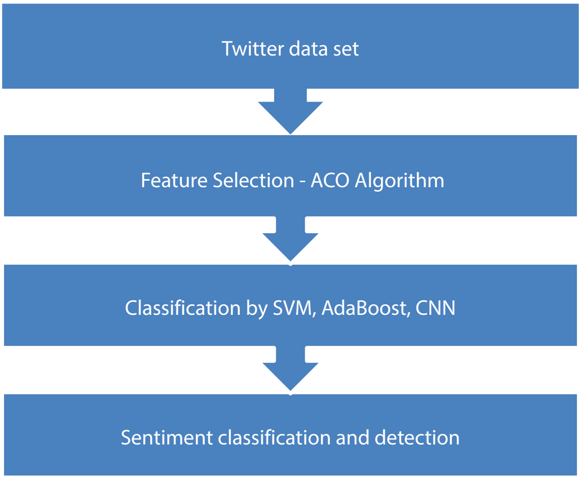

Twitter data set
Feature Selection - ACO Algorithm
Classification by SVM, AdaBoost, CNN
Sentiment classification and detection

Figure 9.1 ACO-CNN deep learning model for sentiment classification and detection.

pixels. For an object to be accurately classified, it is required to first extract the distinctive qualities of the item from an image. It is possible to extract features from photos by employing approaches that have been tried and tested for a long time at feature extraction. For the purpose of determining whether or not an image displays pedestrians, a number of binary classifiers, such as support vector machines (SVMs), are utilized. To accomplish this, the support vector machine (SVM) constructs a hyperplane that connects the non-linear data points of the images. After that, it classifies each point into one of two binary categories. By increasing the distance between the data point and the hyperplane, it is possible to get a higher level of classification accuracy. One example of its application is the procedure of employing a support vector machine (SVM) to identify favorable characteristics in photographs [18].

AdaBoost is the name given to the innovative gradient-boosting approach that was first developed by researchers at the University of Michigan for the purpose of binary classification. After the initial tree decision tree has been constructed, its performance on the training data is utilized to assess the usefulness of the tree. This evaluation will establish whether or not the tree decision tree is useful in your decision-making process. Through the utilization of a number of distinct classification techniques, this approach results in the generation of a single strategy that encompasses everything. The initial model is constructed by beginning with the data required for training. Other models are built to correct any errors that may have been present in the initial model. If they are able to accurately predict the data contained in the training set, then the models are considered to have been developed; otherwise, they continue to be created until the maximum number of models possible is reached. Ultimately, the process of integrating all of the earlier classification models resulted in the creation of the most effective classification model. As a result of its ability to detect pedestrians, the AdaBoost sensor has garnered a lot of attention from professionals. To begin the process of computing the feature values, we must begin by translating the images into what are known as rectangular frames. Simply by designating the series of windows in any order, pedestrians and other people who use the road can be easily differentiated from one another. The same steps are taken as before, but this time the windows in the example image are picked in a different order than they were in the previous instance. With the exception of that, there has been no alteration. The procedure of categorizing windows can be repeated until a cascade of categorization criteria is constructed [19]. Any windows that are not rejected by any of the models are regarded to be pedestrian windows. There is no limit to the number of times you can perform this operation.

When confronted with this circumstance, the first model disregards windows that are obviously not belonging to pedestrians, the second model disregards windows that are less obviously not belonging to pedestrians, and so on.

On the other hand, when applied to big datasets, machine learning techniques have very little practical application. There are a number of reasons for this, including underfitting, sophisticated models, and inadequate optimization of resources. The effectiveness of the procedures is immediately blamed on these impediments, which are directly accountable for the reduction. By applying deep learning networks to big datasets, individuals may be able to acquire new knowledge, improve their ability to predict outcomes, and take action based on those predictions. The implementation of "deep learning" has made it possible for computer models to acquire the capability of learning from both textual and visual data. In response to the boom in the amount of data that is available, a number of distinct deep learning architectures have come into being. The effectiveness of these designs has been demonstrated to be superior when compared to other conventional machine learning techniques.

In the realm of deep neural networks, a convolutional neural network is among the most popular and extensively utilized designs. The area of computer vision makes use of this type of network (CNN). Layers that comprise a convolutional neural network (CNN) include the convolutional layer, pooling layer, activation layer, and connected convolutional layer. A deep convolutional neural network (CNN) is characterized by an extensive network of interconnected convolutional layers (Figure 9.1). Information traveling through the convolutional layer is primarily filtered by it. In the convolutional layer, there is a filter that is applied to a small subset of the input picture's pixels, say 3 x 3.

An operation known as "dot" is applied to the pixel values by the filter, and the effect of this operation is controlled by a weight that has been calculated in advance. Because of this, the size of the picture's data point matrices is decreased once the convolutional layer has been applied to the image. Through the process of back propagation with the matrices that are transmitted to the activation layer, the matrices of the activation layer train the network and provide the network with nonlinearity. The size of the filter matrix and the number of sample layers are both reduced by the process of pooling, which involves merging samples. The sort of layer known as a max layer selects only one property from each category rather than all of them. The output of the max layers is utilized by the linked layer to come up with a collection of candidate label probabilities. To obtain these probabilities, the connected layer is utilized. It is important to take into account the most likely consequence while selecting a label.

9.4 Conclusion

The objective of behavior analysis, also known as sentiment analysis (SA), is to determine the emotional reaction of individuals to a particular service, business, or product. When it comes to identifying and extracting subjective information, such as views and attitudes, from a given language, sentiment analysis makes use of a wide variety of methodologies, strategies, and technologies. Throughout its history, SA has made it a priority to evaluate any and all information that is available to the general public on the internet. At the same time as it is utilized in combination with topic analysis, sentiment analysis has the capacity to accurately detect both positive and negative feelings. The capacity of Deep Learning Techniques to teach both supervised and unsupervised categories makes them an invaluable tool for Sentiment Analysis, which allows for the analysis of sentiment. DLTs are comprised of a number of different types of networks, including feature presentation, phrase modeling, text creation, word representation estimation, vector representation, and sentence classification. This chapter presents a deep learning-based technique to perform sentiment analysis on Twitter scrapped data.

References

1. Bollen, J., Mao, H., Zeng, X., Twitter mood predicts the stock market. *J. Comput. Sci.*, 2, 3, 1–8, 2011.
2. Karabulut, Y., Can Facebook predict stock market activity?, SSRN eLibrary, pp. 1–58, 2013, http://ssrn.com/abstract=2017099 or http://dx.doi.org/10.2139/ssrn.2017099. Accessed 2 Feb 2014.
3. Liu, B., Sentiment Analysis and Subjectivity, in: *Handbook of Natural Language Processing*, Marcel Dekker, Inc., New York, NY, USA, 2009.
4. Richmond, J.A., Spies in ancient Greece. *Greece Rome (Second Series)*, 45, 01, 1–18, 1998.
5. Rawat, R., Kaur, U., Khan, S.P., Sikarwar, R., Sankaran, K., *Using Computational Intelligence for the Dark Web and Illicit Behavior Detection*. IGI Global. 1, 2022, https://doi.org/10.4018/978-1-6684-6444-1.
6. Kharde, V.A. and Sonawane, S.S., Sentiment Analysis of Twitter Data: A Survey of Techniques. *Int. J. Comput. Appl.*, 139, 975–8887, 2016, doi: 10.5120/ijca2016908625.

7. Khairnar, J. and Kinikar, M., Machine learning algorithms for opinion mining and sentiment classification. *Int. J. Sci. Res. Publ.*, 3, 6, 1–6, 2013.
8. Zhang, T., An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. A review, Association for the Advancement of Artificial Intelligence (www.aaai.org), 2001.
9. Vateekul, P. and Koomsubha, T., A Study of Sentiment Analysis Using Deep Learning Techniques on Thai Twitter Data, 2016.
10. Pang, B. and Lee, L., Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.*, 2, 1–2, 1–135, 2008.
11. Chauhan, D., Singh, C., Rawat, R., Dhawan, M., Evaluating the Performance of Conversational AI Tools: A Comparative Analysis. *Conversational Artificial Intelligence*, pp. 385–409, 2024.
12. Castellanos, M., Dayal, M., Hsu, M., Ghosh, R., Dekhil, M., U LCI: A Social Channel Analysis Platform for Live Customer Intelligence, in: *Proceedings of the 2011 international Conference on Management of Data*, 2011.
13. Zhang, W., Yoshida, T., Tang, X., Ho, T.B., Improving effectiveness of mutual information for substantival multiword expression extraction. *Expert Syst. Appl.*, 36, 10919–10930, 2009, doi: 10.1016/j.eswa.2009.02.026.
14. Liu, B., Sentiment Analysis and Opinion Mining. *Synth. Lect. Hum. Lang. Technol.*, 5, 1–167, 2012, doi: 10.2200/S00416ED1V01Y201204HLT016.
15. Yu, Y., Lin, H., Meng, J., Zhao, Z., Visual and Textual Sentiment Analysis of a Microblog Using Deep Convolutional Neural Networks. *Algorithms*, 9, 41, 2016, doi: 10.3390/a9020041.
16. Hagenau, M., Liebmann, M., Neumann, D., Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decis. Support Syst.*, 55, 685–697, 2013, doi: 10.1016/j.dss.2013.02.006.
17. Kim Boes, A.I., Buhalis, D., Wilkinson, P.F., An Empirical Study on the Relationship between Twitter Sentiment and Influence in the Tourism Domain. *Ann. Tour. Res.*, 28, 1070–1072, 2012, doi: 10.1016/S0160-7383(01)00012-3.
18. Pak, A. and Paroubek, P., Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *Proc. Seventh Conf. Int. Lang. Resour. Eval.*, pp. 1320–1326, 2010, doi: 10.1371/journal.pone.0026624.
19. Rawat, R., Chakrawarti, R.K., Sarangi, S.K., Choudhary, R., Gadwal, A.S., and Bhardwaj, V., eds. *Robotic Process Automation*. John Wiley & Sons, 2023.
20. Patsariya, M. and Rajavat, A., A Progressive Design of MANET Security Protocol for Reliable and Secure Communication. *Int. J. Intell. Syst. Appl. Eng.*, *12*, 9s, 190–204, 2024.
21. Rathi, M. and Rajavat, A., Investigations and Design of Privacy-Preserving Data Mining Technique for Secure Data Publishing. *Int. J. Intell. Syst. Appl. Eng.*, *11*, 9s, 351–367, 2023.
22. Dubey, P. and Rajavat, A., Effective K-means clustering algorithm for efficient data mining, in: *2023 2nd International Conference on Vision Towards*

Emerging Trends in Communication and Networking Technologies (ViTECoN), pp. 1–6, IEEE, 2023, May.

23. Nahar, S., Pithawa, D., Bhardwaj, V., Rawat, R., Rawat, A., Pachlasiya, K., Quantum Technology for Military Applications, in: *Quantum Computing in Cybersecurity*, pp. 313–334, 2023.
24. Pithawa, D., Nahar, S., Bhardwaj, V., Rawat, R., Dronawat, R., Rawat, A., Quantum Computing Technological Design Along with Its Dark Side, in: *Quantum Computing in Cybersecurity*, pp. 295–312, 2023.
25. Rawat, R., Telang, S., William, P., Kaur, U., C.U., O., *Dark Web Pattern Recognition and Crime Analysis Using Machine Intelligence*. IGI Global, 2, 2022, https://doi.org/10.4018/978-1-6684-3942-5.

10

Opinion Mining Using Classification Techniques on Electronic Media Data

Meenakshi

Apeejay Stya University Sohna, Haryana, India

Abstract

Both businesses and consumers have changed their viewpoint in reaction to the movement of the Web from being a producer of information to being a receiver of information. Consequently, an increasing number of individuals are opting to make judgments using online platforms. Companies highly value these evaluations as they provide impartial depictions of the customer's genuine sentiments. Although possessing such information ensures substantial advantages, it also presents notable challenges.

One significant factor behind this phenomenon is the widespread availability of social media applications. Prior to the advent of technology, individuals relied on personal networks of family and friends to obtain information for informed purchasing choices.

However, in the present day, the internet has taken control, and individuals rely on online assessments to inform their purchasing decisions. Moreover, rather than engaging in face-to-face interactions, numerous individuals seek for these types of evaluations on the internet. A plausible rationale for this predicament is the surge in the prevalence of social media. Internet users now have the ability to access a vast amount of information that may be advantageous to both companies and their customers. Although this data is undeniably beneficial, the overwhelming number of views required for just one product has resulted in choice fatigue. To mitigate choice fatigue for individuals and facilitate effective decision-making for enterprises, the root-cause analysis approach offers a comprehensive and accurate understanding of products. Aspect extraction, often referred to as opinion target extraction, and sentiment identification, also called opinion mining,

Email: mt6458@gmail.com

Rajesh Kumar Chakrawarti, Ranjana Sikarwar, Sanjaya Kumar Sarangi, Samson Arun Raj Albert Raj, Shweta Gupta, Krishnan Sakthidasan Sankaran and Romil Rawat (eds.) Natural Language Processing for Software Engineering, (155–168) © 2025 Scrivener Publishing LLC

are two components of root-cause analysis. Sentiment analysis is to ascertain the reviewer's overall position about a certain product. Consequently, both customers and companies may more effectively evaluate the advantages and disadvantages of their products. Both organizations and consumers can gain advantages by enhancing the root-cause analysis method proposed in this thesis. Support Vector Machines (SVM), Random Forest, Multilayer perception and CNN algorithms are efficient in opinion mining task.

Keywords**:** Opinion mining, machine learning, electronic media analysis, natural language processing

10.1 Introduction

Electronic media analysis refers to the process of examining and comprehending media content that is transmitted through electronic platforms, such as the internet, radio, and social media. Electronic media analysis aims to decode the information transmitted through various channels and understand its effects on society and individuals. There are other approaches to examining electronic media, including:

Content analysis: This involves thoroughly analyzing the language, graphics, and themes of a program or website, which are the components of media messaging. Through the examination of media information, content analysts are able to identify patterns, such as the occurrence rate of a specific sort of message or the way in which a given issue is presented.

Discourse analysis: To accomplish this, we need to analyze the terminology and speech used by the media to identify the extent to which it affects our perspective of the world. Through the process of discourse analysis, the assumptions and values that are included within media messages may be retrieved from them.

Audience analysis: Studying the cognitive and emotional responses of different demographic groups to media messages is a crucial aspect of this discipline. Different groups of people have different reactions and understandings of media messages, and these understandings affect their attitudes and actions. Conducting audience analysis can assist in identifying these distinctions.

Social media analysis: As part of this procedure, data from social media platforms are examined to provide further insights as to how users communicate and share information on these websites. By utilizing social media analytics, we are able to discern patterns in the data shared on these platforms and analyze the manner in which individuals engage with it.

Natural language processing (NLP) is a subfield of artificial intelligence that focuses on comprehending how humans and computers communicate with one another. It has the potential to be a useful instrument for analyzing information that is presented in the form of text, audio, and video. The scope of this concept encompasses a broad variety of techniques that enable robots to comprehend and generate material that may be utilized by humans. Natural language processing and programming languages are two quite different things. There is a huge distinction between the two. On the other hand, the collection of mathematical operations that they have made available is not static. A computer program that instructs a machine on what to perform is what creates the definition of a computational task. There are currently no compilers or interpreters available for natural languages such as Hindi, English, German, and French. This is a very unfortunate situation. When it comes to language processing, the process often involves converting human-spoken utterances into numerical data, which computers may subsequently use for analysis and interpretation. In addition to this, its objective is to produce natural language writing that is semantically consistent and intelligible. Natural language generation (NLG) and natural language understanding (NLU) are both acronyms that stand for "natural language generation" and "natural language understanding," respectively. A sufficient amount of emphasis cannot be placed on the intricacy of human language. While this is going on, natural language processing (NLP) is emerging as a solution to the growing demand for computer programs that are able to comprehend human pronunciation [1].

Natural language processing techniques that were used in the past depended on linguistics-based strategies, such as part-of-speech (POS) tagging, which drew a solution from the most fundamental semantic and syntactic aspects of a language. An increasing number of people are using NLP into their day-to-day lives. For example, when a user begins typing a search phrase into a search engine, an application can attempt to estimate the following words that the user will type. Smartphones, tablets, laptops, and even ships that already have software loaded on them attempt to recognize human spoken orders and attempt to carry them out. Numerous well-known services, such as chatbots, social media monitoring, language translation, voice assistants, and spam filters, are dependent on the natural language reading capabilities of computers. Natural language processing (NLP) has witnessed the introduction of a broad variety of unique solutions in recent years, which are employed by a wide spectrum of customers. These solutions have been developed by different companies. Among these users are those who construct websites, do research, create conversational artificial intelligence applications, and a great many more [2].

10.2 Opinion Mining

The fast rise of Web 2.0 has led to a significant amount of content being shared on electronic media platforms such as reviews, comments, tweets, blogs, and more. Businesses can gain valuable insights by examining this sort of data. The analysis of electronic media material was previously ignored. In the present day, several firms heavily depend on content analysis to ascertain the public's perception of their brand. However, commercial organizations now face a significant hurdle in accurately gauging public sentiment from internet sources. In this context, there is a need for a completely self-governing system that can gather, assess, and categorize user sentiment among a vast amount of digital media data. Sentiment analysis (SA) is the field of computer science that focuses on analyzing public opinion, attitude, and judgement using subjective input such as text, audio, video, and more. Opinion Mining is an area of social learning.

Social learning refers to the acquisition of knowledge and skills via active engagement and collaboration with others. It involves the exchange of information and knowledge among individuals or groups, usually facilitated by digital platforms. Classrooms, online communities, social media groups, and professional networks are all instances of informal and formal settings where social learning may occur. Students engage in social learning through verbal communication, collaborative effort, and mutual knowledge acquisition. By engaging in their studies, students have the opportunity to enhance their communication, critical thinking, and problem-solving skills, as well as develop a deeper understanding of the subject matter [3].

Furthermore, individuals can gain advantages in both their personal and professional spheres by developing and nurturing social interactions and networks through social learning. Emotions play a vital role in driving behavior, highlighting the importance of social learning in sentiment analysis. The firm use social learning to analyze and monitor the content trends associated with the product. Consumer satisfaction significantly influences repurchase behavior and word-of-mouth recommendations. Individuals have the ability to exercise their voting rights to either support or oppose a leader depending on their sentiments towards his beliefs and personality. Authorities possess the power to implement necessary actions to prevent potentially violent demonstrations or marches if public opinion towards these activities suggests a high probability of violence.

Opinion mining utilizes machine learning, data analysis, and natural language processing techniques to derive quantitative sentiment assessments from unstructured text. There are several approaches available for doing this task. Practically, one may manually do the task by conducting online searches for authentic blog posts, thoroughly reading them, and thereafter assessing the sentiment or polarity of the comments area. However, an algorithm can accomplish the task with more efficiency and precision. Data-driven sentiment analysis systems undergo a process of iterative evolution as machine learning algorithms are used to perform sentiment scoring. Sufficient training samples, in the form of annotated texts tagged with sentiment, are essential for training sentiment analysis algorithms. Experts extract and annotate these samples of training data [4].

10.3 Related Work

Recent deep learning efforts have generated state-of-the-art models that have sparked interest in sentiment analysis. The main difficulty in past sentiment analysis research is understanding the semantic relationships between target properties of sentences in their respective contexts. The reason for this is the significant correlation between the contextual factors of the input phrases and the extent to which they match the desired attribute. Therefore, it is advisable to include both contextual and targeted training connections. Authors developed a model that is reliant on the target. The model employs two Long Short-Term Memory (LSTM) units, positioned on both the right and left sides, to get representations that are contingent upon the target. Furthermore, it implicitly extracts the semantic connection between the goal feature of the text and the contextual factors in the ASA task. This approach can augment traditional word embeddings by including target aspect word representations. An Attention Mechanism is employed to capture the relationship between the aspect and its context [5].

The authors developed a flexible deep learning system using Apache Spark for analyzing massive amounts of data on smartphones. The findings indicate that DL with spark has superior efficiency in comparison to other spark models.

The authors offered two more strategies that may be used to extract relevant attitudes from the surrounding contexts of targets. Additionally, they applied spectral clustering (SC) to the industry-standard dataset following the training of the model using a typical backpropagation technique. Compared to the baseline models, this one had superior performance in properly forecasting emotion. Several research have demonstrated that deep learning algorithms have the ability to automatically collect important input from the ASC. Nevertheless, most of these approaches evaluate the target data directly [6].

The authors introduced two LSTM algorithms that focus on extracting target and aspect information using attention. Therefore, the attention mechanism in the SC paradigm enhances the ASC task by assigning greater significance to the most crucial components of a sentence. The experimental results indicated that both techniques offer a superior understanding of the connections between context and sentiment, in comparison to the baseline models. In their study, the authors showed that utilizing a RAM with attention mechanism may efficiently gather pertinent information from different words in a phrase. They also proposed a technique to measure the significance of each word and its context by employing a multiple attention mechanism [7].

The authors introduced a unique approach to predict sentiment polarity. This strategy depends on understanding the temporal connections between different aspects. The authors created a novel target-based methodology to tackle this issue. This model depicts the relationship between each word and its corresponding context word inside the aspect. This model employs a Convolutional Neural Network (CNN) to address the limitations of the traditional attention mechanism in identifying specific features. The TSC target information was included into word representations by deep transformation. Consequently, the model consistently outperforms several advanced models in terms of accuracy and acquires a deeper understanding of abstract contextual information [8].

The authors emphasized the significance of independently and collaboratively obtaining background and target information. The interactive attention network model is able to comprehend the interactions between target and contextual information by isolating them. This model had impressive performance on the SemEval 2014 datasets, showcasing its ability to acquire significant features for context-based and target-based Aspect Sentiment Analysis (ASA). The attention mechanism in the model enables it to identify the target features that hold the most significance by analyzing their semantic orientation [9].

To preserve the associations between aspect words and their surrounding context, a model based on attention-based encoders was developed. In their study, the authors employed a positional attention mechanism with a penalized feature to improve the performance of the ASA task inside a single sentence. The objective was to enhance the disparity of attention weights closer to the aspect [10].

The ASC problem was addressed by analyzing Chinese review sentences using Knowledge-Enhanced Neural Networks, as described. The objective was to ascertain subjective and situation-dependent components. The sentiment knowledge graph was utilized to ascertain the sentiment polarity of every aspect-pair. The model demonstrated superior performance compared to conventional models when evaluated on the Chinese automobile review dataset, resulting in more thorough sentiment analysis outputs [11].

The authors constructed a deep memory network for sentiment classification by employing a location encoding and content attention strategy. Most SA studies uncover explicit attitudes, whether they are voiced explicitly or indirectly. Explicit emotions have gained considerable attention and made great progress in both the corporate and academic domains. The absence of vocabulary to articulate certain emotions indicates that unspoken sentiments continue to be a problem [12].

To address the issues related to implicit sentiment analysis (SA), the authors suggested employing Bi-LSTM with multi-polarity orthogonal attention. The Bi-LSTM model is more effective than the usual attention mechanism in identifying important aspects with sentiment information in implicit sentiment analysis. Aspect-embedding has been extensively employed for aspect group categorization in ASA activities. However, the depiction of the relationship between aspect-terms and aspect-categories is insufficient [13].

The authors suggested a position-based hierarchical technique to extract important information from phrase segments. The results obtained from trials done on four standard datasets demonstrated that the proposed strategy surpassed other strategies in terms of improving ASC performance. The authors developed a novel approach that incorporates LSTM into the hidden representations to enable both coarse and fine-attention techniques. Rather than depending on location data, most deep learning approaches emphasize the importance of attention when representing each component. ASC's performance may be enhanced by leveraging the multi-level knowledge from SSC and incorporating the position information from several stages. To tackle the problem of long-range dependencies, the authors introduced a novel attention-based model that captures the semantic connections among the identified features in a phrase. The IMDB and Yelp 2018 datasets were extensively studied to enhance the performance of sentiment classification by acquiring conceptual semantic understanding of documents [14–18].

10.4 Opinion Mining Techniques

The progression of digital services in the domain of electronic media has been extraordinary. Microblogging encompasses personal perspectives, firsthand encounters, and recommendations, making it the most effective medium for sharing daily events. This multimedia material includes textual information, photographs, and hyperlinks to external websites. Social marketing, political campaigns, academics, and journalists heavily depend on electronic media to distribute their perspectives on current events and commercial products. The internet greatly streamlines our everyday life by enabling the efficient completion of complex tasks. Online ratings and reviews have become essential in consumers' decision-making process when buying things. Customer feedback is a vital component of product development. Electronic media, the latest digital platform, allows people to share the most current local, national, and global news. An effective method for understanding people's viewpoints on various matters is by utilizing sentiment analysis (SA), which relies on online customer reviews. The information produced by electronic media users effectively conveys customers' thoughts, emotions, and demands on infrastructure, products, and services.

10.4.1 Naïve Bayes

The Naive Bayes approach is a probabilistic classification technique that applies the Bayes theorem. This method of data classification assumes that characteristics are independent of one other, meaning that the presence of one characteristic does not restrict the presence of another feature. The phrase "naive" accurately characterizes it. Despite this oversimplification, Naive Bayes often produces good outcomes in practical situations, especially in the context of text classification problems. This technique utilizes the probabilities of the attributes in the data point to ascertain the probability of the data point belonging to a specific class. Firstly, we calculate the prior probability of each class by dividing the training data by the total number of observations in that class. Next, when presented with a class, we calculate the conditional probability of each characteristic. The chance of a data point belonging to a specific class is obtained by multiplying the prior and conditional probabilities of all its attributes. However, when the independence requirement is adequately robust, Naive Bayes may be used to other classification problems. These approaches provide benefits because of their computational efficiency and ability to achieve excellent performance with less training data.

10.4.2 Support Vector Machine

Support Vector Machine (SVM) is a supervised technique commonly employed for regression and classification tasks. The procedure ascertains the optimal hyperplane for segregating the data points into their respective classes. The support vectors are the closest data points to each class, and the hyperplane is chosen to optimize the margin. Support vector machines (SVM) are commonly utilized in the classification of pictures, texts, and bioinformatics. In high-dimensional data, when the number of attributes exceeds the number of data points, the data become even more valuable. SVM is well-known for its ability to handle non-linearly separable data. It accomplishes this by utilizing kernel methods to convert the input into a feature space with higher dimensions. The goal of support vector machines (SVM) is to find the hyperplane that maximizes the margin while minimizing classification error. SVM does this by solving a constrained optimization problem, which involves finding the Lagrange multipliers that satisfy the Karush-Kuhn-Tucker (KKT) conditions. By resolving this optimization problem, we derive the optimal hyperplane for separating the data points into their respective classes, along with the decision function that utilizes this hyperplane to classify newly additional data points. The determination of the decision boundary for the support vector machine is contingent upon the specific issue and model being used.

10.4.3 Decision Tree

Two instances in which decision trees are utilized are classification and regression. To provide this method its functionality, the feature space is partitioned in a recursive manner. The feature and threshold that effectively split the data into the target classes are selected by the algorithm at each phase of the process. This process is repeated for each and every subset until a certain endpoint is reached, such as a maximum depth for the tree or a minimum number of samples for each node. It is possible to make a prediction about a new class after the creation of the decision tree by tracing the path that the tree takes from its root to its leaf node. Most of the training examples that are able to reach that leaf node are successful in achieving the class. They are a well-liked machine learning method due to their ability to handle a combination of continuous and categorical data, as well as their ease of use and interpretability. The unfortunate reality is that they are not always the most effective, and there is a potential that they will be overfit. There is a possibility that they will not perform as well when dealing with noisy data or a decision boundary that is intricate.

10.4.4 Multiple Linear Regression

Multiple linear regression (MLR) is a statistical approach used to predict a continuous dependent variable YY based on one or more independent variables $YX1$, $XX2$,..., $XXnn$. Due to its inability to directly predict categorical variables, it is rarely used as a classifier. Typical implementations encompass logistic regression, decision trees, support vector machines, neural networks, and random forests, all of which are employed to address classification problems. These approaches are more effective at predicting categorical outcomes. Converting the category variable into a numerical one allows for the utilization of multiple linear regression as a classifier. The multiple linear regression model can utilize modified categorical data as an independent variable. However, it is important to note that this system has some imperfections. It may exhibit worse performance compared to other categorization methods or encounter difficulties when dealing with data that have a high number of categories. Therefore, it is rarely employed as the main answer for a classification problem.

10.4.5 Multilayer Perceptron

In the field of machine learning, the multilayer perceptron (MLP) is a well-known feedforward artificial neural network that is utilized for supervised learning tasks such as classification and regression. The input layer, the hidden layer (or layers), and the output layer are the three components that make up the architecture of a multi-layer perceptron (MLP). It is at the input layer that the model obtains the attributes that it uses as input. Every single node in the input layer is responsible for representing a single input feature. The processing takes place in the hidden layers, which are composed of a collection of nodes that apply non-linear modifications to the data that are being received. The nodes that make up a hidden layer collect data from all of the layers that are below them, assign a weight to the inputs, apply an activation function that is not linear to the sum, and then ultimately transfer the output to the nodes that are below it. A possible output value is represented by each node in the output layer, which is where the model's final output is formed once it has been processed. Before training a multi-layer perceptron (MLP), it is necessary to adjust the weights of the connections that exist between the nodes of the network. This is done with the intention of reducing the disparity between the actual outputs and the predicted outputs. In most cases, the optimization technique known as stochastic gradient descent is utilized for this purpose. The authors are able to maintain the accuracy of the weights by repeatedly updating the weights at each layer of the network, beginning with the output and working their way backwards. Backpropagation is the phrase that accurately describes this process.

10.4.6 Convolutional Neural Network

Convolutional Neural Networks (CNNs) are utilized in various text categorization applications (Severyn & Moschitti, n.d.). Examples of text analysis tasks include subject categorization, sentiment analysis, and spam detection. A word embedding matrix is a widely used format for input data in text classification. In this matrix, each row corresponds to a word vector, while each column corresponds to a vocabulary word. Word embeddings are fed into a 1D convolutional layer, which applies filters to the input sequence, enabling the usage of CNNs for text classification. Typically, the filters have different widths, allowing the network to detect patterns with different levels of detail. Following the processing of the output by the convolutional layer, it is then passed on to a max-pooling layer. This layer selects the highest value from each feature map and utilizes it to reduce the dimensionality of the feature maps. The ultimate outcome, often a probability distribution over the possible categories, is produced by transmitting the resulting feature vector across a fully linked layer. To train the network, the weights are adjusted so that the expected output and the actual labels have minimal discrepancy. Popular techniques for this problem involve the use of gradient descent and backpropagation. Multiple benchmarks have demonstrated that Convolutional Neural Networks (CNNs) attain the highest level of performance, which is promising for solving text classification challenges. The performance of the model relies significantly on the quality and amount of the training data, the selection of hyperparameters, and the model architecture, similar to other machine learning models.

10.4.7 Long Short-Term Memory

The Long Short-Term Memory (LSTM) [19, 20] Network is an architecture of neural networks specifically designed to process sequential input, such as text or voice. Long Short-Term Memory (LSTM) networks [21–23], a kind of Recurrent Neural Network (RNN), have the ability to selectively retain or ignore information from previous time steps. Due to their ability to capture long-term relationships and interactions between inputs, activities that include sequential data are very successful in achieving their goals. An LSTM cell is a unit that has the ability to retain information over an extended period of time by utilizing a sequence of memory cells. The gates on these memory cells enable fine control over the incoming and exiting data streams. The input gate decides the data that should be incorporated into the memory cell at the current time step, while the forget gate determines the data that should be eliminated from the memory cell at the previous time step. Currently, the output gate specifies which specific information from the memory cell should be used to generate the output.

10.5 Conclusion

The complexities inherent in human speech provide significant challenges in identifying and extracting pertinent segments from customer assessments. The strategies given in this chapter are essential for efficiently extracting elements from reviews. The objective of sentiment analysis is to ascertain the overall viewpoint of the reviewer on the product under consideration. Consequently, both firms and customers may have the ability to evaluate the advantages and disadvantages of their products more accurately. By enhancing the proposed method of analyzing the fundamental causes of a problem, both organizations and customers may enjoy the advantages. Regarding opinion mining, algorithms such as CNN, Multilayer Perceptron, Random Forest, and Support Vector Machines (SVM) have strong performance. Enhancing the model with additional guidelines to effectively process sarcasm will be a forthcoming improvement. Data patterns with even minimal occurrences can be detected by employing supplementary criteria. This requires an extremely strong model that is capable of accurately capturing even the smallest variations. The authors will explore the possibility of integrating this model into their design or modifying their architecture to incorporate these guidelines at a later time. Further elimination of outdated data can enhance forecast accuracy, even when reinforcements are employed to capture shifts in trends. The ranking issues offer a comprehensive understanding of the significance of each aspect, with the most crucial ones being the study of time requirements and money considerations. To enhance corporate decision-making, it is advisable to prioritize issues that may be promptly resolved or need low expenditure.

References

1. Patil, A. and Thakare, S., Analyzing public sentiment variations on and Facebook. *Int. J. Adv. Res. Comput. Sci. Technol.*, 5, 2, 30–3, April–June. 2017.
2. Sharma, A., Sharma, A., Singh, R.K., Upadhayay, M.D., Hybrid classifier for sentiment analysis using effective pipelining. *Int. Res. J. Eng. Technol.*, 4, 8, 2276–81, August 2017.
3. Mishra, A. K., Tyagi, A. K., Dananjayan, S., Rajavat, A., Rawat, H., Rawat, A., Revolutionizing Government Operations: The Impact of Artificial Intelligence in Public Administration. in: *Conversational Artificial Intelligence*, pp. 607–634, 2024.

4. Zhang, K., Yu Cheng, Y., Agrawal, A., Palsetia, D., Lee, K., Choudhary, A., SES: sentiment elicitation system for social media data. *Icdmsentire*, pp. 129–36, 2011.
5. Tupsoundarya, A. and Dandannavar, P.S., Sentiment expression via emoticons on socialmedia. *Int. J. Res. Appl. Sci. Eng. Technol.*, 6, 6, 2404–8, June 2018.
6. Sharif, W., Samsudin, N.A., Deris, M.M., Naseem, R., Mushtaq, M.F., Effect of negation in sentiment analysis. *Int. J. Comput. Linguist. Res.*, 8, 47–56, June 2017.
7. Nahar, S., Pithawa, D., Bhardwaj, V., Rawat, R., Rawat, A., Pachlasiya, K., Quantum technology for military applications. in: *Quantum Computing in Cybersecurity*, pp. 313–334, 2023.
8. Perdana, R.S. and Pinandito, A., Combining likes-retweet analysis and naive Bayes classifier within twitter for sentiment analysis. *J. Telecommun. Electron. Comput. Eng.*, 10, 1–8, 41–6, April 2018.
9. UmaMaheswari, S. and Dhenakaran, S.S., Sentiment analysis on social media big data with multiple tweet words. *Int. J. Innov. Technol. Explor. Eng.*, 8, 10, 2514–8, August 2019.
10. Sharma, A. and Dey, S., Artificial neural network based approach for sentiment analysis of opinionated text. *ACM Transaction*, pp. 37–42, 2012.
11. Turney, P., Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *Proc. Assoc. Comput. Ling. (ACL)*, pp. 417–24, 2002.
12. Pang, B. and Lee, L., Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2, 1–2, 1–135, 2008.
13. Tan, S., Li, Y., Sun, H., Guan, Z., Yan, X., Bu, J., Chen, C., He, X., Interpreting the public sentiment variations on twitter. *IEEE Trans. Knowl. Data Eng.*, 26, 5, 1158–70, May 2014.
14. Filho, P.B., Avanço, L., Pardo, T., Nunes, M.D.G.V., NILC–USP: an improved hybrid system for sentiment analysis in Twitter messages, in: *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval2014)*, 2014.
15. Gonçalves, P., Araújo, M., Benevenuto, F., Cha, M., Comparing and combining sentiment analysis methods, 2014.
16. Iqbal, M., Karim, A., Kamiran, F., Bias-aware lexiconbased sentiment analysis, in: *ACM Symposium on Applied Computing*, pp. 845–50, 2015.
17. Vidya, N.A., Fanany, M., II, Budi, I., Twitter sentiment to analyze net brand reputation of mobile phone providers. *Procedia Comput. Sci.*, 72, 519–26, 2015.
18. Sikarwar, R., Shakya, H. K., Kumar, A., Rawat, A., Advanced Security Solutions for Conversational AI. in: *Conversational Artificial Intelligence*, pp. 287–301, 2024.
19. Patsariya, M. and Rajavat, A., A Progressive Design of MANET Security Protocol for Reliable and Secure Communication. *Int. J. Intell. Syst. Appl. Eng.*, 12, 9s, 190–204, 2024.

20. Rathi, M. and Rajavat, A., Investigations and Design of Privacy-Preserving Data Mining Technique for Secure Data Publishing. *Int. J. Intell. Syst. Appl. Eng.*, 11, 9s, 351–367, 2023.
21. Dubey, P. and Rajavat, A., Effective K-means clustering algorithm for efficient data mining, in: *2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN)*, 2023, May, IEEE, pp. 1–6.
22. Chauhan, D., Singh, C., Rawat, R., Chouhan, M., Conversational AI Applications in Ed-Tech Industry: An Analysis of Its Impact and Potential in Education, in: *Conversational Artificial Intelligence*, pp. 411–433, 2024.
23. Chauhan, D., Singh, C., Rawat, R., Dhawan, M., Evaluating the Performance of Conversational AI Tools: A Comparative Analysis, in: *Conversational Artificial Intelligence*, pp. 385–409, 2024.

11

Spam Content Filtering in Online Social Networks

Meenakshi

Apeejay Stya University Sohna, Haryana, India

Abstract

In the contemporary period, individuals engage in both professional and personal contacts mostly through electronic modes of communication. Over the past decade, the popularity and influence of email and social media platforms such as LinkedIn, Facebook, and Twitter have experienced a significant and rapid increase. Online social networking platforms and email are exploited by spammers to distribute their messages, taking advantage of their popularity. The spam filtering system must possess sufficient robustness to effectively detect and prevent unwanted communications, hence halting spammers' activities promptly. When attempting to recognize spam communications, the majority of individuals rely on either text-based or collaborative methods. To improve the accuracy of spam detection in emails and social networks, it is necessary to employ strategies that reduce the occurrence of false positives and false negatives. Support vector machines are the most effective method for categorizing and detecting spam emails.

Keywords**:** Spam e-mail filtering, machine learning, support vector machine, accuracy, online social networks

11.1 Introduction

Electronic communication has become an essential component of contemporary existence. There are two fundamental categories of electronic communication: formal emails and personal online social networks (OSNs). The biggest issue that internet users are currently facing is the widespread

Email: mt6458@gmail.com

Rajesh Kumar Chakrawarti, Ranjana Sikarwar, Sanjaya Kumar Sarangi, Samson Arun Raj Albert Raj, Shweta Gupta, Krishnan Sakthidasan Sankaran and Romil Rawat (eds.) Natural Language Processing for Software Engineering, (169–182) © 2025 Scrivener Publishing LLC

occurrence of spam messages. Spam refers to unsolicited commercial communications delivered over electronic messaging systems with the intention of marketing, spreading hazardous content, engaging in phishing, or just annoying customers. Spam may easily become viral because individuals tend to trust and spread information that others have already shared. Spam is an unproductive and often costly use of time for consumers. Internet consumers encounter the challenging issue of spam, but there are several methods to safeguard themselves from it [1].

11.1.1 E-Mail Spam

In the present era, the vast majority of individuals opt to interact using electronic mail. An urgent issue is the development of a method to effectively exclude undesirable emails from the email system. A spam filter must possess the qualities of being user-friendly and adaptable to be considered necessary. The exact and accurate filter only produces lower rates of false positives and false negatives. False positives arise when a filter evaluates a message and erroneously categorizes it as spam. The erroneous forecast made by the filter results in unsolicited emails being classified as false negatives. An effective spam filter should have a low percentage of false positives and a high rate of accuracy. Due to the potential danger caused by spam emails, categorizing them as "ham emails" in users' inboxes raises the likelihood of unwittingly causing damage to their systems [2].

Another concern is to the elevated frequency of false positives. Although ham email may not be harmful, it nevertheless gets filtered into the spam folder. As a result, the user can overlook important emails due to her failure to remember to check her spam folder. For a classifier, maintaining a low percentage of false positives and false negatives is of utmost importance. A negative connection exists between the rates of true positives and true negatives, and the rates of false positives and false negatives. Consequently, as the rates of false positive and false negative decrease, the accuracy rate increases proportionally with the number of correct classifications. Spam filter attacks may be classified into two primary categories: poison attacks and impersonation attacks. David and his colleagues (2018a) discussed instances of poison attempts in an email conversation. Poisoning is a technique used to decrease the chances of spam detection by including many non-spam terms into spam emails. The spam e-mail perpetrator employs impersonation assaults by hacking into the accounts or computers of unsuspecting users. There are two main approaches to spam filtering: content-based and identity-based. The content-based filter analyzes and ranks emails based on patterns and phrases that are commonly associated with spam. Identity-based spam filtering utilizes a comparison of the sender's email address with a blacklist and a whitelist to accurately detect spam.

Both the whitelist and the blacklist include email addresses of persons whose communications should not be subjected to filtering by the spam filter [3].

11.2 E-Mail Spam Identification Methods

Various techniques can be employed to avoid spam emails, including the utilization of intricate mathematical algorithms to ascertain if an email was solicited or unsolicited. These materials exhibit a significant level of spatial and temporal intricacy. To overcome these issues, there are other alternatives available that do not require any technological expertise. Below is an incomplete compilation of a few of them:

11.2.1 Content-Based Spam Identification Method

Through the utilization of a static spam term corpus, the content-based e-mail spam detection system achieves its goal. With the purpose of categorizing the emails into spam and ham folders based on the frequency with which particular spam keywords appear in the emails respectively. Spammers are able to circumvent content-based spam detection technologies by changing and filtering their emails before sending them to the people they plan to send them to. There is a pattern that spammers may follow while they are spreading spam [4].

As part of their analysis of spam emails, researchers provided a comprehensive breakdown of the methods that spammers employ. The rule-based filtering technique is included in the content-based spam detection strategies. For the purposes of training and email classification, rules are utilized. When it comes to the identification and filtering of spam e-mails, users are required to manually build pattern matching rules, which can be a process that can be time-consuming. Since the activity of spam emails changes over time, it is necessary to have a rule that changes to recognize spam emails. As an illustration of a rule-based algorithm, the RIPPER rule learning algorithm is a good example.

One of the strategies that falls under the area of content-based methods is the identification of spam messages using statistical methods. For the purpose of modeling the data, approaches from the field of machine learning are utilized. Several algorithms have been developed for the purpose of classifying texts; however, for these approaches to be effective, the classifier must first be trained by a human. Among the various types of content-based searching, there is a subgroup that makes use of search engines. Individuals that engage in spamming send emails that contain connections to other websites. The link will lead you to a website that is sponsored by advertisements. It is the responsibility of the search engine to automatically review the content of the websites that are linked in the emails.

11.2.2 Identity-Based Spam Identification Method

The identification of the senders is crucial for the detection of spam emails. This identification mechanism utilizes a list-based method, which incorporates whitelists and blacklists. Users using electronic mail have the ability to create a whitelist and blacklist. Users have the ability to add email addresses to a whitelist or blacklist. Messages originating from email addresses listed in the whitelist are authorized to be delivered to the intended recipients' primary email folders, but messages originating from email addresses listed in the blacklist are automatically directed to the spam folder. In a distributed adaptive blacklist situation, the server is tasked with the responsibility of managing the adaptive blacklist. Upon receiving an email, the e-mail transfer agent triggers the activation of the filter. Examining the email's header may be used to determine whether mails are spam, serving as a method for categorizing spam based on identity. This approach verifies if emails lack a sender ID in the "From" field and if the "To" field has an excessive number of recipient IDs.

11.3 Online Social Network Spam

Email is not the only medium in which spammers and spam messages may be found; social media platforms are also plagued by them. The growth of online social networks (OSNs) such as Facebook, LinkedIn, Twitter, and others is accelerating at an unparalleled rate. It is possible that the unanticipated growth of online social networks (OSNs) might be attributable to two factors: the large number of users and the greater connectedness to the general public. Ignoring the bad features of social networks is not going to get you very far, despite the fact that there are clearly some beneficial advancements in the domain of social networks currently. The most aggravating experience for people who use social media is having to cope with an excessive amount of spam. The amount of inaccurate or unpleasant information that is moving throughout the networks will lead to the emergence of a number of issues that will affect society. Screening out undesired texts may be accomplished in a variety of different ways. The number of spam messages that are transmitted through social media platforms has not greatly diminished, despite the fact that this has been stated [5].

Different types of social spam are available. Common among them are of following types:

- **Bulk Message Spam**

Disseminating several identical copies of material to numerous social network users all at once. Bulk texting may also be used to send ads, false messages, or harmful links to social networks.

- **Malicious Link Spam**

To deceive users, harm their devices, or generate revenue for spammers, malicious links are posted on social networks. Users' comments or postings are a perfect medium for the harmful links to spread.

- **Fake Review Spam**

Users write false reviews to benefit financially from the product service suppliers that reply to them. Its secondary objective is to lessen the product's appeal.

- **Click-bait Spam**

Simply said, clickbait is an enticing headline that, when clicked, takes the reader to a website that offers irrelevant or boring information. In this way, clickbait boosts their financial gain by increasing the amount of page views. Like jacking is a sort of clickbaiting when people post status updates on social media without any prior knowledge or intention.

11.4 Related Work

Spam is a term that refers to unwanted or unwelcome messages that are sent over communication channels such as instant messaging, microblogging sites, and electronic mail. When it comes to email spam filters, users and email service providers (ESPs) alike have a lot of things to complain about. It is necessary for filters to be both resistant to attacks and extremely accurate for them to efficiently identify spam.

The addition of more spam filters ought to result in a reduction in the number of false positives and false negatives. A false positive is the term used to describe the situation in which legitimate emails are mistakenly selected as spam. It is referred to as a false negative when emails that are considered spam are not identified as spam. In spite of the vast number of spam classifiers that are currently available, it is still possible for some spam filter attacks to occur. These attacks include the poison attack and the impersonation attack. A number of popular words are used into poison assaults to reduce the likelihood that spam e-mails would be recognized as spam. Spammers engage in impersonation assaults, in which they give the impression that they are genuine users by utilizing stolen credentials or by compromising their computers in some other way. With this in mind, the objective is to recognize spam and develop a technique of spam filtering that can efficiently differentiate between legitimate emails and those that are not wanted [6].

The email system was improved by adding data from social networks to better boost the effectiveness of spam screening. It was found in a number of studies that when it came to filtering spam, individuals' trust, hobbies, and the degree of closeness in their relationships on social networks were also taken into consideration. However, due to the fact that spam emails are totally generated from social networks, these factors are not sufficient to identify and classify their content. For the purpose of classifying the emails, there were no criteria that were specifically based on the email itself that were calculated. An active learning-based spam message categorization was proposed by the authors as a means of significantly reducing the amount of time required for classification while still preserving accuracy. The technique that they adopted did not take into consideration the categorization of emails that contain photographs.

The utilization of a fuzzy-based technique was the concept behind the classification of spam emails. After conducting an analysis of the structural patterns of the spam messages, the authors decided to implement a fuzzy-based method to reduce the noise point problem. The authors indicated that the selection of features for spam filtering in emails is based on category ratio and word frequency, and that this decision is made independently of the sample size obtained from each class.

One of the potential solutions that the authors have suggested is to use information gathered from social networks to identify spam in emails that were received. The spam filtering strategies that it utilized included both identity-based and content-based methods, in addition to social network approaches. For the purpose of scanning and grading emails that fall under the content-based category, typical spam keywords and patterns are utilized. In the content-based category, impersonation attacks, in which hostile actors appear to be genuine users by compromising their computers or forging their IDs, are extremely frequent. These assaults are considered to be particularly common. One of the most important aspects of identity-based filtering is the whitelist and blacklist of email addresses that are managed by the user. Due to the fact that they rely only on sender email addresses, they are more susceptible to poison assaults. A form of attack that involves cramming spam emails with actual phrases in an effort to make them less likely to be detected as spam is called a spear phishing attack. Email systems rely on a subset of social network indicators, such as proximity, interest, and trust, to differentiate between genuine and spam communications. This allows them to discern between the two types of messages. Bayesian spam filters were utilized to significantly improve accuracy, provide protection against attacks, and ensure effective spam detection. In addition, trust and interest criteria were not taken into account while calculating email networks [7].

Methods of machine learning were utilized by the authors to identify individuals that engage in spamming on Twitter. A significant portion of the information that was utilized for the empirical research consisted of millions of tweets. It was determined that a neighborhood-based detection method would be the most effective way to identify spammers using Twitter. The collection of data on profile-based feature evasion and the validation of evasion tactics were utilized to detect spammers on Twitter.

We were able to detect spam accounts on Twitter by looking at criteria such as the number of followers and tweets that a user had. A thorough investigation was conducted on the 24 detecting factors as well as the many evasion tactics that are utilized by spammers on Twitter. However, the difficulty is that the datasets that were gathered from Twitter and categorized as harmless only contained people who had never posted malicious URLs. This is the case even though some of the accounts that were included in the datasets might possibly be hazardous [8].

Researchers have devised a technique that may be used to identify spammers on social media platforms. For the purpose of identifying spammers, a method that is based on supervised machine learning was proposed. The contents of the message and the actions taken by the user were related with a number of essential features, and the authors used some of these qualities. For the purpose of identifying spammers, feature selection strategies were utilized in conjunction with Support Vector Machine (SVM) classification algorithms. This allowed for the determination of which qualities were the most important and the amount of weight that should be assigned to each feature. In addition, statistical analysis and human selection were utilized in the process of automatic feature extraction.

The method obtained a high level of performance when it took into consideration the true positive rate of spammers as well as the rate of non-spammers. The technique was proposed by scientists as a means of tracking down potentially hazardous emails [9].

This method, which placed a focus on persistent threats, was superior than other existing tactics because of its recipient-oriented aspect. Preprocessing of the emails with data that was particular to the firm was performed before a random forest classifier was used to categorize the emails that corresponded to the company. Following that, certain properties were retrieved from the mixture. This is very harmful. E-mail is delivered in small batches to individuals or small groups with the intention of deceiving them into believing that it is harmless and that it is suited to their particular need. The threat actors are identified by the network protector based on the significance and quality of their actions. Standard computer network attacks often target web servers and other network-based listening services as their primary targets. However, targeted attacks frequently employ the strategy of social engineering through the use of email. Regrettably, the majority of communities continue to allow email to reach their social networks, which poses a significant threat to their security [10].

The researchers made a presentation on attentive learning for the classification of emails. Email categorization has lately been a topic of discussion because of the large number of emails that are sent and received on a daily basis. Through the utilization of an attentive learning technique, automatic e-mail categorization is able to imitate the natural and ever-changing behaviors of users. Researchers investigated a number of different email classification methods; their findings were based on the activities of users. This email categorization was made feasible by expressing the structure and content of each incoming message as a subset of the feature space, which is referred to as a feature set. Subsequently, a selection of the traits was selected for additional examination. To restate, the feature sets were utilized to choose set classifiers or processes that would be responsible for determining the correctness of the features that been selected. This selection results in high-accuracy judgments on the classification of e-mails into safe and spam categories, and its behavior is so tight-fisted that it is efficient in terms of the amount of time it takes to complete the task.

It has been recommended that adaptable learning frameworks should be utilized in conjunction with adaptive approaches. Sequential input is received by the learning system that is contained inside the framework. After the forecast has been made, the next step is to incorporate new data into the model that is already in place. The Adaptive Bayesian Classifier with Dynamic Features (ABC-DYNF) provides a control and adaptation mechanism that allows it to quickly react to changes in features as well as drifting that is either actual or contextual. Additionally, the ABC-DYNF investigated the effectiveness of email classification performed with a Bayesian Filter (BF), in addition to analyzing the performance of various pre-processing, control, and adaptive algorithms. Email foldering poses a big difficulty because of the behavior that it exhibits, which is always changing. There is also the possibility of adding, removing, or creating a large number of directories throughout the course of time. It is necessary to pick particular qualities to engage in data streaming for textual data. In the realm of email classifiers, adaptive approaches are a game-changer that could not be ignored [11].

A number of scholars raised objections to the digital investigation.

Within the scope of the digital investigation, the examination of hundreds of thousands of artifacts was carried out. Using email is the sole method that allows for the creation of a substantial volume of data and information. E-mail forensics has been used to a wide variety of instances, and all of these investigations have been based on genuine cases that have been investigated. Not only may criminals use spam email to contact potential victims, but they can also utilize it as a tool to develop their scams. As a result of the fact that no forensic tools are capable of reliably identifying suspicious spam e-mails, digital investigators will need to equip themselves with strong knowledge and patience to discover how the essential data are being transmitted through content analysis. The mission of those working in the field of email spam detection is to reduce the number of false positives and false negatives as much as possible, rather than to attempt to extract useful information from spam emails. Spam sent via email provides criminals with a viable channel through which they may communicate their thoughts and views. Furthermore, spam filtering has been utilized as a tool in the investigation of homicides and other types of criminal activity, despite the fact that the information that is retrieved from garbage emails may not be of any use for digital investigations.

11.5 Challenges in the Spam Message Identification

Unfortunately, content-based spam filtering is more vulnerable to poison attacks, which is the main problem with spam message filtering. This approach involves adding several real terms to spam emails to make them less likely to be identified as spam. The impersonation assault is a problem for the identity-based spam filtering. Common users' identities are used in this assault by hacking their computers or faking their IDs. We need spam filters that can withstand attacks and produce fewer false positives and false negatives. Legitimate emails that are incorrectly flagged as spam are known as false positives, whereas undetected spam emails are known as false negatives. While previous research has taken into account trust, interest, and closeness as social network metrics, it has done so only using OSN data and the premise that a lower hop count indicates a closer relationship. The degree of intimacy between two users is, however, heavily influenced by the nature of their connection. So, for active email filtering, it's important to think about the relationship strength as well. If a person wants to build a decent spam classifier, he/she needs to take into account characteristics from both social networks and the email network while assessing spam emails. The most common formats for spam e-mails include both plain text and pictures. Therefore, a spam classifier is considered effective if it takes that possibility into account as well. The Open Source Network (OSN) already has several spam detection and filtering mechanisms available, however such systems need the user to actively enable a spam filter. Unfortunately, users usually won't go to the trouble of manually setting each rule, thus the system might not operate in real time and might only work for one social network. For efficient spam message categorization, picking the right features is just as crucial as picking the right classifiers.

11.6 Spam Classification with SVM Filter

Within the realm of support vector classification, LIBSVM is an integrated piece of software that is utilized. There is a sophisticated classification method known as support vector machines (SVM), which begins with the separation of the available data into training data and testing data. A supervised learning model known as the support vector machine (SVM) [11, 12] is utilized for classification in addition to regression analysis. On the basis of the training examples, support vector machine (SVM) training models construct a model that assigns the test data that does not have a class label to a category that has been predetermined. There is a known class label included in every training data set, as well as various properties or features of the variables that are being observed. The LIBSM application programming interfaces (APIs) [13–15] are utilized to generate a model that is derived from the training data. This model is then utilized to forecast the values of the test data by providing just the test data properties.

For the purpose of spam filtering [16, 17], the tokens or keywords that are included in the emails, as well as the frequency with which these tokens appear in both the spam and the ham emails, are taken into consideration. A value of one is assigned to ham, whereas a value of zero is assigned to spam e-mails. Both the data used for training and those used for testing are scaled to a certain range. The process of scaling ensures that no token will be able to dominate any other token. In the event that the data is not scaled, some keywords may have a larger frequency, which increases the likelihood that they would dominate the words that appear less often. Furthermore, the existence of the least dominating terms is not taken into consideration at all. Following the scaling of the data used for testing and training, the train APIs are utilized to generate the models. The prediction API is utilized to construct predictions for the testing data, and the accuracy of these forecasts is also computed. These predictions are based on the models that were generated and the scaled test data. The support vector machine (SVM) is able to discover a solution to the optimization equation and the parameters that are necessary for performing this optimization after being provided with a training set of label pairs. During the process of generating the models by utilizing the training APIs, equations are discovered.

In high-dimensional space, the support vector machine (SVM) locates a linear separation hyperplane that has the maximum margin. The support vector machine (SVM) is capable of doing non-linear classifications by utilizing kernel techniques in addition to linear classification. Mapping the data into a space with greater dimensions is the method that is used to do non-linear classification. There are four distinct kernels that may be utilized, and they are the linear, sigmoid, polynomial, and radial basis functions. The Radial Basis is utilized by LIBSVM.

11.7 Conclusion

Despite the existence of many real-time spam message filtering algorithms, users still encounter the issue of receiving spam messages in their email and on online social networks (OSN). Spam messages not only have the potential to irritate customers, but they may also lead to significant consequences such as financial losses and a deterioration of trust among users. Hence, both users and service providers want a dependable system that can rapidly identify and eliminate spam messages. If you want to categorize and identify spam emails, support vector machines are the most reliable option. To enhance the categorization of emails in future research, one will have to collect information on the preferences of users, in addition to the existing attributes, and take into consideration the dynamic and evolving nature of these preferences. Regarding the filtering of undesired messages in online social networks (OSN), future goals involve utilizing supplementary datasets, considering new factors, and evaluating the system's effectiveness using various support vector machine (SVM) kernels. One further possibility under consideration is the development of a browser extension that has the capability to analyze incoming data and effectively eliminate spam messages in the online social network (OSN). This extension would utilize social context variables from each social network to achieve its filtering functionality. This extension would have the capability to function on many networks for the purpose of spam filtering, rather than being limited to just one network.

References

1. Cranor, L.F. and LaMacchia, B.A., Spam! *Commun. ACM*, 41, 8, 74–83, 1998.
2. Jain, G. and Sharma, M., Social Media: A Review, in: *Information Systems Design and Intelligent Applications*, pp. 387–395, Springer, India, 2016.
3. Nelson, M., *Spam Control: Problems and Opportunities*, pp. 23–82, Ferris Research, India, USA, 2003.
4. Cormack, G.V., Hidalgo, J.M.G., Sánz, E.P., Feature engineering for mobile (SMS) spam filtering, in: *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, US, pp. 871–872, 2007.
5. Davidson, T., Warmsley, D., Macy, M., Weber, I., Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, pp. 512–515, 2017.
6. Dadvar, M. and Trieschnigg, D., Towards context-aware filtering of hate speech content on Twitter. *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, pp. 573–576, 2013.
7. Waseem, Z. and Hovy, D., Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. *Proceedings of the NAACL Student Research Workshop*, pp. 88–93, 2016.
8. Ribeiro, M.H., Araújo, M., Gonçalves, P., Benevenuto, F., SentiBench—a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Sci.*, 7, 1, 45, 2018.
9. Zubiaga, A. and Ji, H., Tweet, but verify: epistemic study of information verification on Twitter. *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, pp. 641–650, 2014.
10. Mathew, B.K., Bhattacharyya, P., Ghosh, S., Overview of multimedia data processing and analytics in social media: Extraction, summarization, and content filtering. *Multimed. Tools Appl.*, 78, 2, 1573–1611, 2019.
11. Rathore, N. and Rajavat, A., Smart Farming Based on IOT-Edge Computing: Applying Machine Learning Models For Disease And Irrigation Water Requirement Prediction In Potato Crop Using Containerized Microservices, in: *Precision Agriculture for Sustainability*, pp. 399–424, Apple Academic Press, USA, 2024.

12. Patsariya, M. and Rajavat, A., A Progressive Design of MANET Security Protocol for Reliable and Secure Communication. *Int. J. Intell. Syst. Appl. Eng.*, *12*, 9s, 190–204, 2024.
13. Rathi, M. and Rajavat, A., Investigations and Design of Privacy-Preserving Data Mining Technique for Secure Data Publishing. *Int. J. Intell. Syst. Appl. Eng.*, *11*, 9s, 351–367, 2023.
14. Dubey, P. and Rajavat, A., Effective K-means clustering algorithm for efficient data mining, in: *2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN)*, pp. 1–6, IEEE, 2023, May.
15. Chauhan, D., Singh, C., Rawat, R., Chouhan, M., Conversational AI Applications in Ed-Tech Industry: An Analysis of Its Impact and Potential in Education, in: *Conversational Artificial Intelligence*, pp. 411–433, 2024.
16. Chauhan, D., Singh, C., Rawat, R., Dhawan, M., Evaluating the Performance of Conversational AI Tools: A Comparative Analysis, in: *Conversational Artificial Intelligence*, pp. 385–409, 2024.
17. Park, M. and Fink, E., Examining the effectiveness of social media content moderation: Evidence from Reddit. *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*, pp. 442–453, 2019.t

12

An Investigation of Various Techniques to Improve Cyber Security

Shoaib Mohammad[1*], Ramendra Pratap Singh[1], Rajiv Kumar[1], Kshitij Kumar Rai[1], Arti Sharma[1] and Saloni Rathore[2]

[1]School of Law, IMS Unison University, Dehradun, India
[2]Invertis University, Bareilly, India

Abstract

Cyber security threats, characterized by a series of assault stages, persistently aim to accomplish a pre-established goal. Due to the intricate nature of these attacks, the intruder is capable of bypassing the target's security defenses and gaining access to a majority of its systems. When accessing data kept in the cloud, there is a genuine risk of experiencing data breaches, compromised credentials, Denial of Service (DoS) assaults, hacked interfaces and Application Programming Interfaces (APIs), permanent data loss, and other significant cybersecurity concerns. Due to the constant innovation of cybercriminals, who continuously develop more advanced methods to avoid detection, it is challenging to both identify and prevent these malicious activities. As digital technology advances, gigabytes and terabytes of data are now generated every second. Businesses in a variety of industries are finding that using the internet to manage their resources and transactions is useful. Given the value of data and the need to safeguard its security and privacy, securing big data remains a major challenge for all solutions. Due to the exponential expansion of network data, intrusion detection is becoming increasingly important, and manual analysis would be either impossible or take the same amount of time as analyzing it. As a result, there is an urgent need for an automated system capable of extracting relevant information from enormous amounts of hitherto untapped data when it comes to network intrusion detection. Data mining can perform a variety of tasks, such as clustering, prediction, classification, and the extraction of association rules between data pieces. This paper discusses machine learning techniques for designing intrusion detection systems for big data networks. In this

**Corresponding author*: shoaib.mohammad@iuu.ac

Rajesh Kumar Chakrawarti, Ranjana Sikarwar, Sanjaya Kumar Sarangi, Samson Arun Raj Albert Raj, Shweta Gupta, Krishnan Sakthidasan Sankaran and Romil Rawat (eds.) Natural Language Processing for Software Engineering, (183–192) © 2025 Scrivener Publishing LLC

approach, the NSL KDD data set is used as input. First, the CFS-correlation feature selection approach is used to pick only relevant features from the NSL KDD data set. The NSL KDD data collection contains 41 features. The number of characteristics was reduced to 16 after applying the CFS algorithm. The 16 attributes are then used by machine learning techniques to classify and predict malware data in the NSL KDD data set.

***Keywords*:** Cyber security, attacks, privacy, vulnerability, machine learning, intrusion detection

12.1 Introduction

The Internet is a global network of linked computers that offers companies and people unprecedented opportunities for communication and cooperation. This phenomenon has prompted people to modify their lifestyle, as seen by the increasing prevalence of online education and e-commerce, which are gradually replacing traditional schools and physical retail stores. The vast and highly intelligent World Wide Web (WWW), with its immense volume of data, has had a significant impact on the overall worldwide security situation over a long period of time. Despite the disclosure of many data breaches by leading security firms, users' accounts are still being consistently compromised. Technological advancement is a crucial factor in all of these recent events. This provides an opportunity for criminals to experiment with novel methods of cybercrime, such as exploiting computer systems and networks [1].

Cybersecurity threats [14], including several phases of attack, persistently aim to accomplish a pre-established goal. Due to the intricate nature of these attacks, the intruder is capable of bypassing the target's security defenses and gaining access to the majority of its systems. Accessing data held in the cloud today poses a genuine risk of data breaches, compromised credentials, Denial of Service (DoS) assaults, hacked interfaces and Application Programming Interfaces (APIs), irreversible data loss, and other significant cybersecurity concerns. Due to the constant enhancement of attackers' sophisticated plans and methods to evade detection, it is challenging to detect and prevent these activities. Due to the fact that these assaults mostly target corporate data and other sensitive information, stopping them poses a significant challenge. Attacks use advanced strategies and often include extensive groups of synchronized personnel using technological infrastructure with extensive Command and Control (C&C) networks of computers. High device densities in wireless networks, caused by the rapid rise of the Internet of Things and vehicle networks, lead to data traffic overflow and pose significant security risks. Assailants consider a range of obstacles and possibilities [2].

Cybercrime encompasses a series of deliberate attacks that are carried out in a systematic manner with the objective of achieving a certain purpose. Assailants meticulously adhere to this sequential procedure to successfully execute their attacks. A Defense architecture that is driven by intelligence and incorporates the cyber kill chain provides a series of steps for the identification and prevention of cyber attacks. The typical cyber death chains include seven distinct stages: reconnaissance, weaponization, delivery, exploitation, installation, command and control, and operations on target [3]. During the reconnaissance phase, attackers do research to identify a suitable target who would enable them to accomplish their objectives. Before proceeding to the next step, referred to as "delivery," the weaponization phase will merge the exploit and malware into a payload. For instance, during the third stage of a cyber attack, hackers could use an email deception to transmit malicious software to the target. The term "Zero-Day" refers to the specific code used in the fourth stage of an attack, enabling adversaries to exploit a vulnerability and gain unauthorized access to the victim's machine [4]. Subsequently, in step five, the attackers construct a durable backdoor or insert malicious code into the target website to ensure continuous access for an extended period. In the sixth step, a command network is established, enabling the attacker to exert remote control over the victim using a command and control (C&C) server. This is facilitated by a malevolent application. Ultimately, the intruders successfully achieve the mission's goal by targeting the victims. The intruder has the potential to engage in clandestine surveillance, pilfer credentials, manipulate or interrupt data, sabotage computer hardware, collect and extract information, as well as covertly modify or extract data. Disrupting just one link in the cyber death chain may bring the onslaught to a total stop. Interrupting the chain early will minimize the amount of damage caused [5].

12.2 Various Attacks [6–9]

Phishing is the biggest cyber danger right now. The neologism "phishing" is a play on the word "fishing" that arose from the idea of utilizing bait to lure in victims. One of the most common ways that hackers target networks and individuals is via phishing attacks, which have claimed the lives of a large number of users. It seems to be coming from a trusted source, but it's really a fake link. It is a sting operation set up by an enemy to steal people's private information. It is a malicious link that, if clicked, will take the user to other malicious websites created by the hacker. Spear phishing assaults occur when the phishing website is primarily designed to target a specific person or business. Cybercriminals prey on people's trusting nature to breach security measures; unfortunately, the majority of internet users nowadays fall prey to such tricks whenever they visit a URL. As a kind of social engineering often used to acquire user data, phishing is among the most subtle yet potent assaults. The perpetrators of phishing attacks may quickly concoct and send out phony emails that include harmful code or links. By installing malicious malware on the victim's PC, the attackers were able to steal sensitive information and send it back to themselves via a backdoor when the user clicked on the misleading URL. In this way, the attacker's server will get the victim's request instead of the legitimate ones. Phishing is the primary vector through which most assaults have gained access to systems. Unless they incur massive losses, victims will remain oblivious to phishing attempts. These phishing links have been the primary vector for most users' falls prey. Phishing emails were formerly a common way for cybercriminals to steal personal information.

In a cryptojacking attack, the hacker takes control of a victim's computer, mobile phone, tablet, or other connected home device and uses it to mine cryptocurrency. To implant crypto mining code on the victim's system, malicious onscreen characters use a malicious connection or infect websites or online ads with JavaScript code. One study by Adguard found that in-program cryptojacking was evolving at a pace of 31%. Cryptojacking is gaining popularity because it allows thieves to make a lot of money with very little effort. The majority of one million registration devices were compromised in February 2018 by the Sominru crypto mining botnet. The admins had mined 8,900 Monero, which is worth about 3.6 million dollars, according to Proofpoint, a cybersecurity company. Two Chinese programmers were accused by the US government a year ago of stealing data from 45 US tech companies, government agencies, and more than 100,000 US Navy personnel. According to the indictment, the defendants compromised systems that remotely managed the information technology infrastructure of governments and businesses worldwide. Businesses that store data for other companies on their servers or remotely manage their clients' IT infrastructure are particularly vulnerable to these types of assaults. Hackers may also target consumers by infiltrating these companies' systems. Looking at the development of sophisticated technologies, most companies will choose cloud solutions to host their IT infrastructure. Companies should go with large cloud providers like Google and Amazon since they are much more secure than smaller companies who are becoming more wary of these kinds of assaults. Since it provides transparency and security via encryption, blockchain is being used in many applications. Among the most promising applications of Blockchain technology, smart contracts stand out. Projects like this use blockchain technology to conduct code, and if certain conditions are satisfied, the result is an advanced resource transfer. Smart contracts will be crucial for all of these applications, from completing financial trades to approved innovation security. Experts promise that clever contracts still contain faults, despite the fact that they have possible used cases. Hackers exploited a flaw in Parity, a multi-signature digital wallet, in 2017 and stole $32 million in advance. Due to the inherent simplicity of the blockchain, the primary concern with protecting the privacy of the savvy contract data is its accessibility. The aggressors have acknowledged this vulnerability formally. In 2019, businesses that employ smart contracts will face a huge challenge.

This year, cybercriminals will target mobile phones more than ever before. There was a 54% increase in new mobile malware variants in 2017, according to the Symantec Internet Security Threat Report. The methods and tools used by cybercriminals to monitor iOS and Android devices are constantly evolving. Talos, a security and intelligence company, discovered a hacker group targeting a small number of iPhones in India at the Mobile Device Management (MDM) conference a year ago. It came out that the hackers sneaked into the devices using social engineering attempts and physical access to steal data. Avoiding downloading files from unknown sites and tapping on connections are the best ways to protect yourself against adaptable malware. Also, keep your mobile operating systems up-to-date; this will make them more resistant to digital threats.

"Associations know the advantages of AI innovation to guard their framework, but they are likewise mindful that assailants have interesting capacities to misuse their framework with that same technology," said Rodney Joffe, senior VP of Neustar, in a published statement. To identify digital threats, several cybersecurity resistance businesses have started to use AI models. Despite this, hackers are able to get through these defenses and launch more sophisticated assaults.

Using generative antagonistic systems, in which two neutral systems compete to uncover the AI computations used by the other, is one way to do this. The developers may easily create a model to circumvent the computation if they happen to stumble into it. Programmers may also get into data sets used to train AI models, which poses a threat in and of itself. They may inject malicious code or change markings so threats are marked as safe instead of questionable.

The goal of cyber espionage is to gain an advantage over a targeted corporation or government agency by means of coordinated, sensitive data or protected technology. For Merriam-Webster, "undercover work" is "the act of spying or utilizing spies to acquire data about the plans and exercises particularly of an outside government or a contending organization." Take it online, and the secret operatives are hordes of evil programmers from all over the globe who employ cyberwarfare to further their own political, economic, or military agendas. From government frameworks to financial frameworks or utilities assets, these hand-picked and highly renowned hackers can shut them down with ease. They have impacted the outcome of worldwide events, impacted political contests, and contributed to the success or failure of enterprises. To sneak into systems or frameworks and remain undiscovered for a long period, many of these attackers use advanced persistent threats (APTs). International espionage is not a new phenomenon; it dates back to the early Middle Ages.

The world of clandestine agents has progressed with the times, and we are now facing a more formidable challenge with digital monitoring. This novel kind of planned and organized risk use digital warfare techniques to gain economic, military, or political benefits. Cybercriminals with exceptional skill are recruited with the goal of disrupting or destroying government or military infrastructure, or to gain unauthorized access to financial systems. From altering the outcome of significant political choices to causing a stir at global events, they are quite capable of creating a situation of total chaos on a global scale. The frustrating part about digital surveillance is that the perpetrator usually follows the standard procedure to ensure that their footprints remain undetectable for a considerable amount of time. As this immoral practice has been gradually adopted, the danger has now spread to other sectors as well. The Verizon 2018 Data Breach Investigations Report states that the major initiatives dealing with the problem of digital covert operations are education, assembly, and open organization. To get access to sensitive material belonging to a government agency or a military foundation, cyber surveillance involves the use of intermediate servers and an unauthorized system or framework. Attacks on traded-off information, which often includes sensitive financial, political, or military data, are a well-known example of this notorious behavior. Onscreen characters allied with governments or nation states seeking to get an advantage are the most common risk entertainers in the world of digital monitoring. Recently, FireEye, a cybersecurity company located in California, exposed the goals of APT39, an Iranian cyber intelligence collecting. The gathering's primary focus continues to be the Middle Eastern media transmission companies. It turned out that the group's covert operations were motivated by a genuine fear for Iran.

When one nation utilizes digital weapons like viruses and hackers to impair another's critical computer systems, they are engaging in cyberwarfare. The goal of these assaults is to cause harm, death, or devastation. In future conflicts, hackers will use computer code to assault enemy infrastructure, fighting side by side with regular soldiers utilizing weaponry like firearms and missiles. A shadowy realm populated by spies, hackers, and top-secret digital weapon projects, cyberwarfare is a perilous and ever-present aspect of global wars. But at this moment, there is a genuine danger that things may swiftly spiral out of control due to the absence of clear regulations controlling online combat and the increasing cyberwarfare weapons race. When one country's computers or data networks are targeted by another, either intentionally or unintentionally, by means such as virus attacks or denial-of-service attacks, this is known as cyberwarfare. Decision-makers in the military and civilian sectors may find advice on how to protect their nation's technological infrastructure against the destructive effects of cyberwarfare in RAND studies.

The term "cyber terrorism" refers to coordinated, coordinated, and targeted electronic assaults against a particular target over the Internet or other networks, emanating from a variety of fear mongering sources with a variety of motivations. Typically, those who oppress people's minds online see their goals as integral components of a nation's fundamental structures or as commercial endeavors.

12.3 Methods

This section presents the design of an intrusion detection system for massive data networks using machine learning methodologies. This framework utilizes the NSL KDD data set as its input. Initially, the NSL KDD data set undergoes processing using the CFS-correlation feature selection technique to only extract the most relevant features. The NSL KDD dataset consists of 41 characteristics. The CFS algorithm was used to restrict the number of features to 16. Machine learning algorithms are then used to classify and predict malware data in the NSL KDD dataset based on these sixteen attributes.

CFS, an integral component of feature selection evaluations, determines the value of an attribute based on its distinct capacity and the extent to which it overlaps with other features. We choose attribute subsets based on their high correlation with the class and low correlation with each other [10].

This approach is ideal for solving classification and regression problems. Support Vector Machines (SVM) identify a hyperplane, which is a line in the case of two-dimensional data, that separates the training data into different classes. This knowledge is then used to categorize new data points. The discovery of the hyperplane, which maximizes the distance between classes, enhances the potential for generalizing hidden data. The Support Vector Machine (SVM) algorithm is known for its superior classification performance, since it provides the highest accuracy on the training set. Data overflow [15–18] is not the result of it.

LS Support vector machines [19–21], such as SVM (Least Square), make no assumptions about the data. There are two primary classifications of support vector machines: linear and non-linear. The representation of training data is often done in a linear manner using a line, sometimes referred to as a hyperplane [11].

The theorem itself is a fundamental principle of Bayesian classification. By using a large dataset, Naive Bayesian Classification techniques provide fundamental foundations that are comparable to decision tree and neural network classification. Naive Bayes classification may be used to represent a subset of dependent attributes. This procedure is used to ascertain the posterior probability P(x|c) for each class. In this study, the outcome is forecasted based on the class with the highest probability. Naive Bayes use a similar methodology to predict various probabilities based on different characteristics [12]. If the weight of each instance in the dataset is determined by the previous results of the base classifier for each of these instances, then the AdaBoost Decision tree [13] uses a dataset that is weighted based on these outcomes. If the instance is misclassified, the weight of subsequent models will be increased, whereas it will remain constant if it is properly classified. The final conclusion is obtained using a weighted vote of the basic classification, where the weight of the model is determined by the misjudgment rate.

12.4 Conclusion

Intrusion detection is becoming increasingly crucial as network data continues to grow at an exponential rate, and manual analysis would either be impossible or take the same amount of time as manually analyzing the data would take. Because of this, when it comes to network intrusion detection, there is an urgent need for an automated system that is capable of extracting important information from large amounts of data that has hitherto been untouched. Data mining can be used to perform a wide range of tasks, including clustering, prediction, classification, and the extraction of association rules between data pieces, among other things. The purpose of this study is to examine machine learning techniques for constructing intrusion detection systems for massively parallel networks to prevent cyber attacks.

References

1. Tang, M., Alazab, M., Luo, Y., Big data for cybersecurity: vulnerability disclosure trends and dependencies. *IEEE Trans. Big Data*, 5, 3, 317–329, 2019.
2. Raghuvanshi, A., Singh, U., Sajja, G., Pallathadka, H., Asenso, E., Kamal, M. *et al.*, Intrusion Detection Using Machine Learning for Risk Mitigation in IoT-Enabled Smart Irrigation in Smart Farming. *J. Food Qual.*, 2022, 1–8, 2022, doi: 10.1155/2022/3955514.
3. Vasan, D., Alazab, M., Venkatraman, S., Akram, J., Qin, Z., MTHAEL: cross-architecture IoT malware detection based on neural network advanced ensemble learning. *IEEE Trans. Comput.*, 69, 11, 1654–1667, 2020.
4. Almoman, O., A Feature Selection Model for Network Intrusion Detection System Based on PSO, GWO, FFA and GA Algorithms. *Symmetry*, 2, 1046, 2020, doi: 10.3390/sym12061046.
5. Raghuvanshi, A., Singh, U., Joshi, C., A Review of Various Security and Privacy Innovations for IoT Applications in Healthcare, in: *Advanced Healthcare Systems*, pp. 43–58, 2022, doi: 10.1002/9781119769293.ch4.
6. Davis, J.J. and Clark, A.J., Data preprocessing for anomaly based network intrusion detection: A review. *Comput. Secur.*, 30, 6–7, 353–375, 2011.
7. Reddy, G.T. *et al.*, Analysis of dimensionality reduction techniques on big data. *IEEE Access*, 8, 54776–54788, 2020.
8. Zhang, Y. and Zhao, Z., Fetal state assessment based on cardiotocography parameters using PCA and AdaBoost, in: *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pp. 1–6, 2017.
9. Abdi, H. and Williams, L.J., Principal component analysis. *Wiley Interdiscip. Rev. Comput. Stat.*, 2, 4, 433–459, 2010.
10. Aggarwal, M., Performance analysis of different feature selection methods in intrusion detection. *Int. J. Sci. Technol. Res.*, 2, 6, 225–231, 2013.
11. Kabir, M.E. and Hu, J., A statistical framework for intrusion detection system. *2014 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, pp. 941–946, 2014, doi: 10.1109/FSKD.2014.6980966.
12. Shen, Z., Zhang, Y., Chen, W., A Bayesian Classification Intrusion Detection Method Based on the Fusion of PCA and LDA. *Secur. Commun. Netw.*, 2019, 1–11, 2019, Available: 10.1155/2019/6346708.

13. Shahraki, A., Abbasi, M., Haugen, Ø., Boosting algorithms for network intrusion detection: A comparative evaluation of Real AdaBoost, Gentle AdaBoost and Modest AdaBoost. *Eng. Appl. Artif. Intell.*, 94, 103770, 2020, doi: 10.1016/j.engappai.2020.103770.
14. Revathi, S. and Malathi, D.A., A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection. *Int. J. Eng. Res. Technol. (IJERT)*, 2, 12, 1848–1853, December 2013.
15. Chirgaiya, S. and Rajavat, A., Tiny object detection model based on competitive multi-layer neural network (TOD-CMLNN). *Intell. Syst. Appl.*, 18, September 2022, 200217, 2023.
16. Rathi, M. and Rajavat, A., *Analysing Cryptographic and Random Data Sanitization Techniques in Privacy Preserving Data Mining*, vol. 83, Allied Publishers, New Delhi, India, 2023.
17. Rathore, N. and Rajavat, A., Scalable edge computing environment based on the containerized microservices and minikube. *Int. J. Softw. Sci. Comput. Intell. (IJSSCI)*, *14*, 1, 1–14, 2022.
18. Dhar, S., Dhar, U., Rajavat, A., Factors and Attributes of Team Players: A Study of Engineering Students in India, in: *International Simulation and Gaming Association Conference*, pp. 53–60, Springer International Publishing, Cham, 2021, September.
19. Pithawa, D., Nahar, S., Bhardwaj, V., Rawat, R., Dronawat, R., Rawat, A., Quantum Computing Technological Design Along with Its Dark Side, in: *Quantum Computing in Cybersecurity*, pp. 295–312, 2023.
20. Shrivastava, M., Patil, R., Bhardwaj, V., Rawat, R., Telang, S., Rawat, A., Quantum Computing and Security Aspects of Attention-Based Visual Question Answering with Long Short-Term Memory, in: *Quantum Computing in Cybersecurity*, pp. 395–412, 2023.
21. Rawat, R., Chakrawarti, R.K., Raj, A.S.A., Mani, G., Chidambarathanu, K., Bhardwaj, R., Association rule learning for threat analysis using traffic analysis and packet filtering approach. *Int. J. Inf. Technol.*, *15*, 6, 3245–3255, 2023.

13

Brain Tumor Classification and Detection Using Machine Learning by Analyzing MRI Images

Chandrima Sinha Roy[1]*, K. Parvathavarthini[2], M. Gomathi[3], Mrunal Pravinkumar Fatangare[4], D. Kishore[5] and Anilkumar Suthar[6]

[1]Department of Computer Science and Engineering, JIS College of Engineering, Kalyani, West Bengal, India
[2]Department of Computer Science and Engineering, Vels Institute of Science, Technology and Advanced Studies, Chennai, India
[3]S. A. Engineering College, Tamil Nadu, India
[4]Computer Science and Engineering, Department of Polytechnic and Skill Development DVK MIT World Peace University, Kothrud, Pune, India
[5]Department of ECE, Aditya College of Engineering & Technology, Surampalem, India
[6]New LJ Institute of Engineering and Technology, Ahmedabad, Gujarat, India

Abstract

In recent years, there has been a rise in the death rate as a consequence of the proliferation of encephaloma tumors among individuals of all ages. The identification of physical tumors is a challenging endeavor that requires a significant amount of time and effort on the part of medical professionals. This is due to the complex nature of the tumors and the presence of unwanted noise in the MR imaging data. Since this is the case, the diagnosis and location of tumors at an early stage are of the highest significance. The use of medical imaging in conjunction with segmentation and relegation techniques has the potential to deliver an accurate and speedy diagnosis. This is accomplished by monitoring and forecasting the progression of cancer at its various stages. In this article, we provide a method that is based on machine learning and is used to segment and categorize magnetic resonance imaging (MRI) pictures for the purpose of detecting brain cancers. To segment pictures, extract features, and categorize them, this system makes use of the

**Corresponding author*: mail2chandrima@gmail.com

Rajesh Kumar Chakrawarti, Ranjana Sikarwar, Sanjaya Kumar Sarangi, Samson Arun Raj Albert Raj, Shweta Gupta, Krishnan Sakthidasan Sankaran and Romil Rawat (eds.) Natural Language Processing for Software Engineering, (193–206) © 2025 Scrivener Publishing LLC

Naive Bayes method, the Support Vector Machine technique, and the K Nearest Neighbor algorithm.

Keywords**:** Brain tumor, classification, machine learning, accuracy, precision, recall

13.1 Introduction

The fatality rate has dramatically increased as a result of the growth in the number of encephaloma tumors that have been seen across all age groups over the course of the last several years. The two primary classifications of tumors that form in the brain in an uncontrollable manner are known as benign and malignant disorders. The term "malignant" is used to describe tumors that have an irregular structure that results in the proliferation of cancerous cells from the tumor. Benign tumors are characterized by the presence of normal structures and cells that do not contribute to the development of cancer. When attempting to detect tumors based on physical criteria, medical practitioners face a challenging and time-consuming task. This is due to the fact that tumors are complex and MR imaging data includes signals that are not intended. In light of this, it is of the utmost importance to promptly diagnose and localize the precise position of the tumor. By integrating medical imaging with segmentation and classification algorithms, it is possible to get an accurate early diagnosis. This may be accomplished by monitoring and predicting cancer-prone regions at various levels. [1] Tags are used to safeguard the material of the user.

The procedure of segmenting MRI images, which is both laborious and time-consuming, is required to identify the tissues that are present in brain tumors. With the assistance of segmentation, the structures in medical pictures that are often misinterpreted may be accurately discovered and diagnosed. In the event where the radiologist manually assessed the tumor by utilizing an excessive number of pictures, it is possible that an inaccurate diagnosis was made. It is necessary to have an automated system in place to guarantee that the processing and interpretation of medical images are free of errors caused by humans. the third one Digital image processing encompasses a wide range of fields, including geology, photography, microscopy, astronomy, computer vision, and medical imaging, among others. The process of doing research in the domains of medicine and science comprises a multitude of intricate phases. Medical imaging is a vital initial step that must be taken before artificial intelligence (AI) [25–27] may begin to segment medical photos and assist in computer-aided diagnosis (CAD) [25, 26, 28]. They streamline the process of treatment planning and execution, which brings about an improvement in the efficiency of human-machine interaction during surgical operations. In order to create diagnostic tools that are very helpful in the area of medicine [29, 30], this method comprises the creation of imaging equipment as well as the execution of a therapeutic strategy. An extensive range of medical technologies were used to generate the images of the human body that are shown here. Computer tomography (CT) and magnetic resonance imaging (MRI) are two of the most important instruments that we have at our disposal for non-invasive imaging of the human body. "[4]" is the text that was entered by the user; "[5]" is the text that represents the user.

The buildup of abnormal tissues [2, 3, 11, 14, 19] inside the brain is the most prevalent mechanism that leads to the development of brain tumors. However, these abnormal tissues disrupt the normal processes of tissue synthesis, growth, and death, which causes them to proliferate and increase in an uncontrollable manner. Several different types of medical imaging techniques, such as computed tomography (CT) and magnetic resonance imaging (MRI), have the potential to detect brain cancer instances. Diagnostic imaging techniques, such as computed tomography (CT) and magnetic resonance imaging (MRI), are commonly used by medical professionals in the process of searching for brain malignancies. By providing radiologists and physicians with three-dimensional (3D) images of the brain, magnetic resonance imaging (MRI) and computed tomography (CT) scans may be used to detect brain cancer. Through computerized analysis, specialists may be able to save time and provide more accurate conclusions. This kind of analysis may swiftly detect aspects of a brain tumor and give a comprehensive three-dimensional image. Through the use of automated or semi-automatic tumor segmentation, professionals have the opportunity to liberate their time from the responsibilities of patient treatment planning and follow-up management as shown in Figure 13.1: Steps Involved in MRI Image Processing. The text that they are using is "[6]".

Several obstacles impede the prompt processing of whole MRI images. Examples of these challenges include the presence of bias fields, noise interference, and the inclusion of non-brain tissues. To address this particular problem, we offer a wide array of pre-processing alternatives at our disposal. After processing, photographs need to go through the time-consuming pre-processing phase, which involves removing any non-essential features. Prior to commencing this procedure, it is important to complete the pre-processing phase [4]. Image grayscale conversion, noise reduction, and reconstruction are sequential stages in the pre-processing

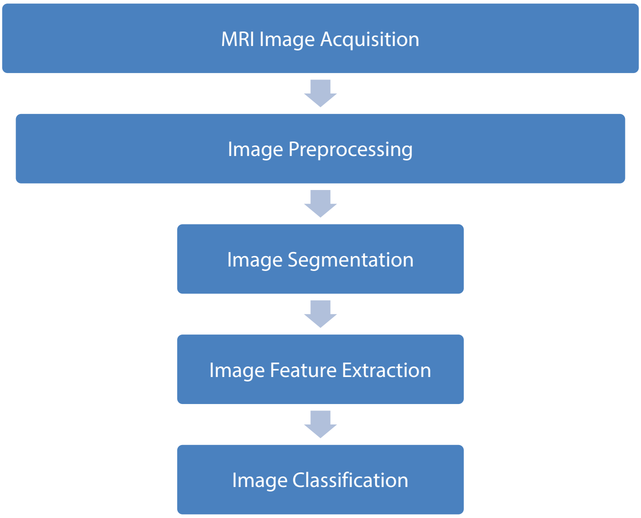


Figure 13.1 Steps involved in MRI image processing.

pipeline. Grayscale image conversion is a commonly used pre-processing method [1]. Various filtering methods are used to eliminate further noise after grayscale processing of the image. To get satisfactory outcomes while obtaining photographs from the database, it is essential to eliminate any background noise. The current approaches being used for noise reduction have notable disadvantages.

Segmentation refers to the process of dividing something into distinct segments or parts. Large images were generated during the process of scanning; medical professionals should possess the ability to visually identify these images manually within a suitable timeframe. It plays a crucial role in clinical diagnostics and is used in computer-assisted surgery and pre-operative planning.

This approach is called feature extraction, and it involves assigning a feature vector to each character simultaneously, which is essentially a representation of the character. Ultimately, the goal is to achieve two challenging objectives: improving the identification rate with the fewest components possible and generating a consistent feature set for several occurrences of the same sign. The current feature extraction approaches failed to prioritize the features for examination in later diagnostics.

Classification: Each piece of data inside a batch is organized into a certain category from a set of predetermined options. Typically, this approach is used to categorize brain images as either normal or malignant. The main objective of classification is to accurately predict the target class for each occurrence in the dataset. The algorithm accomplishes this task by categorizing brain pictures as either containing tumors or being devoid of malignancy. Considering the lack of emphasis on this stage in present approaches, our suggested endeavor would largely concentrate on precisely recognizing MRI images.

13.2 Literature Survey

It is feasible to get information from photographs of the brain by using techniques that are associated with image processing. A number of medical disorders have been successfully diagnosed and treated with the use of magnetic resonance imaging (MRI), which has shown to be effective. Pre-processing a picture consists of the following processes, which are as follows: There are several techniques that may be used to improve the quality of cranium scans, reduce noise, and get rid of artifacts. This image should make it easier to identify the malignancies that are present. To enhance magnetic resonance imaging (MRI) brain imaging of astrocytomas, Suryavamsi *et al.* [7] proposed three techniques: "Histogram Equalization," "Contrast Limited Adaptive Histogram Equalization," and "Brightness Preserving Dynamic Fuzzy Histogram Equalization." The PSNR, RMSE, and MSE metrics are examples of performance measures that have been used to verify these methodologies and quantify the results they provide.

Noise removal from magnetic resonance imaging (MRI) data is the first and most crucial stage in the process of acquiring the target signal. Earlier phases in the pre-processing phase included the use of nuisance regression and independent component analysis. Bianca De Blasi and her colleagues [8] devised a number of different LD cleaning processes to exclude non-BO LD signals from healthy patients as well as those who were suffering from temporal lobe epilepsy. In comparison to data that was only preprocessed in the 0.01-0.1 Hz area, all of the preprocessing pipelines that were investigated improved the temporal features, such as the total signal-to-noise ratio (StNR) and power spectrum density, in the frequency range that corresponds to the resting state. Before beginning the ICA procedure, the authors went through the whole pre-processing pipeline to get the DMN. These pipelines and groups fared better than others when it came to characterizing the posterior cingulate cortex. This was the case when compared to other pipelines and groups.

Pre-processing techniques that have been proposed by Poornachandra and Naveena [9] have the potential to enhance the categorization of glioma tumors, like tumors that are seen in the brain. State-of-the-art deep learning has recently been used in the field of medical imaging. With enhanced segmentation findings, researchers are able to get a better knowledge of brain tumors, which ultimately results in more accurate disease identification and novel therapy choices for patients.

The use of magnetic resonance imaging (MRI) for the diagnosis and staging of cervical cancer has been the subject of intense discussion. The consistency of the picture makes the process of image segmentation much more difficult than it would otherwise be. To enhance the Region Scalable Fitting approach for image segmentation, Setiawan Widyarto and colleagues [10] included pre-processing procedures prior to using a region-based active contour model. In regional areas, the models that were utilized included data on intensity as one of its components. Within the context of the preprocessing step, the 2D-sigmoid function was used to perform processing on the tumor border. The addition of a 2D-sigmoid function during the pre-processing stages resulted in an increase in the contrast of the brain MRI image.

Important characteristics must be extracted by algorithms used for tumor segmentation, a total of eleven. The accuracy of brain tumor segmentation was increased by Jui *et al.* by the use of an enhanced feature extraction component. This component takes into consideration the association between the compression that is induced by the growth of the tumor and the structural deformation experienced inside the brain. MRI volumetric data from the lateral ventricle (LaV) is modified via the use of three-dimensional nonrigid registration and deformation modeling. To validate and make use of LaV deformation feature data for brain tumor segmentation, we will use conventional classification algorithms. Real and virtual patient photographs were used by the author to carry out a comprehensive quantitative and qualitative analysis of the component that was considered for implementation. All indications point to a good outcome for the probe.

Instead of using tissue segmentation or nonlinear registration, Jun Zhang and colleagues depend on landmarks in their suggestion for the extraction of longitudinal structural magnetic resonance images for the purpose of Alzheimer's disease identification. It is necessary for you to locate the distinctive landmarks in the training images before you will be able to recognize them in the testing photographs in a quick and efficient manner. The need for nonlinear registration and tissue segmentation will be avoided by using an approach that identifies landmarks in a short amount of time. With these points of reference in hand, we gather longitudinal data that is relevant to the environment, together with advanced statistical aspects, to evaluate the brain's spatial absorption. In terms of efficiency and performance, the experimental findings demonstrate that our suggested technique outperforms the method that is considered to be the state-of-the-art. This is the case when the method is evaluated on the Alzheimer's Disease Neuroimaging Initiative database, particularly for mental cognitive impairment and Alzheimer's disease. It was determined that the categorizing work was finished with a rate of 88.30 percent [12].

BLADeS is an all-encompassing and completely automated breast computer-aided diagnosis (CAD) system that was created by Gabriele Piantadosi and her colleagues [13]. This system was designed to assist in the identification of breast cancer and lesions. There are a number of modules that Michael Osadebey and his colleagues proposed for a hierarchical design. Some of these modules include breast segmentation, motion artifact reduction, lesion localization, and malignancy-based categorization. For the purpose of providing a fair comparison, the researchers used a cross-validation sample consisting of 42 patients who had histological lesions that were confirmed. The results of the experiments demonstrate that BLADeS was capable of diagnosing breast lesions using T1-weighted DCE-MRI in a totally automated manner, without any involvement from a human being at any stage of the processing chain (processing chain). Tags were used to safeguard the material of the user.

Hsin-Yi Tsai and colleagues [15] presented a novel method that makes use of the Gray-Level Co-Occurrence Matrix (GLCM) to accelerate the process of graphics processing unit (GPU)-based feature extraction. In contrast to its sequential version, which is developed and optimized on a single computer by using MATLAB and C, this technique is deployed on several GPU devices for the purpose of optimizing performance. The suggested technique outperforms the serial version when it is evaluated on GeForce GTX 1080s, with speedups occurring between 25 and 105 times for single-precision MR brain shots and between 15 and 85 times for double-precision MR brain images. For the purpose of testing, the actual percentage of improvement is determined by the size of the ROIs that are employed.

An automatic segmentation approach that is based on Convolutional Neural Networks (CNNs) was reported by Sérgio Pereira *et al.* [16]. More precisely, the system was based on small kernels that were 3 x 3. Not only does reducing the weights of the network provide assistance in preventing overfitting, but it also makes it possible to create patterns that are more detailed. This pre-processing step is not used by CNN-based segmentation algorithms very often; nevertheless, when paired with data augmentation, it yields good results for the segmentation of brain tumors shown in magnetic resonance imaging (MRI) pictures.

The enhanced picture quality that is provided by ultrahigh field (7 T) magnetic resonance imaging (MRI) was used in a semi-automatic segmentation approach that was described by Jinyoung Kim and colleagues [17]. For the purpose of this method, several other structural MRI modalities were used to offer supplemental edge data. The integration of susceptibility-weighted, T2-weighted, and diffusion magnetic resonance imaging (MRI) results in the production of a one-of-a-kind edge indicator function. Each of these modalities was customized to meet the requirements that were special to it. When it comes to geometric active surfaces, having a grasp of the form and structure of the subcortical structures enables more efficient development. To physically separate neighboring buildings from one another, an iterative approach was used. In case that the structures did not overlap, a penalty was given to the division at their borders. There are fifty distinct regions of the brain that are engaged, as stated by Antonios Makropoulos *et al.* [18], beginning with the early preterm period and continuing until the age that is comparable to full term. This study makes use of a unique segmentation approach to investigate the intensity distribution across the whole brain. This is done while taking into consideration the anatomical hierarchy and the physical restrictions. This technique is distinct from the conventional procedures that are based on atlases because, in contrast to manual reference segmentations, it increases the resolution of label overlaps. The findings of the experiments demonstrate that the suggested method is extremely dependable throughout a broad spectrum of gestational ages, ranging from 24 weeks to term-equivalent.

13.3 Methods

Image enhancement and refinement are essential for enhancing its quality. Mobile phone images include several elements that have the ability to affect the segmentation results. The pre-processing of an image involves three stages: scaling, noise reduction, and enhancement. The digital picture may include a multitude of sounds. Basic thresholding is troublesome since it might lead to picture noise generated by ineffective capture. Therefore, it is essential to minimize image noise. Image noise refers to a random alteration in the colors or intensities of a picture. Photographic noise may manifest in several forms, such as periodic noise, film grain, quantization noise, shot noise, salt and pepper noise, and Gaussian noise. To reduce these disruptions, one may use filters such as the Median and Wiener filters. Morphological techniques provide a range of approaches to accomplish noise reduction. Gaussian filtering has the effect of reducing the noise in pictures, but median filtering has the potential to alter the luminosity of individual pixels. Here, the researchers used the GF to eliminate extraneous noise. Instead of using the importance of intensity at each individual pixel, Gaussian filtering used a weighted average of intensities derived from neighboring pixels.

The number 18 was included. Filtering improves the interpretability of pictures for humans and aids in the development of increasingly advanced image processing systems. By manipulating the histogram of the input image to a fixed value, we can guarantee that the distribution of intensities stays unchanged. This technique reliably enhances the overall difference between light and dark areas in photos, which is especially valuable when the image's relevant information is closely related to contrast. By using this technique, it is possible to achieve a uniform distribution of intensities in the histogram. In areas characterized by low contrast, this distinction is significant. Histogram equalization achieves image enhancement by evenly distributing the most common intensities across the picture.

Following the use of dynamic fuzzy histogram equalization to enhance the image, it is then partitioned into many sections and segmented to generate the accurate Region of Interest (ROI). Essential features are then retrieved from these parts. Picture segmentation is the process of identifying and categorizing separate sections within an image. Data may be segmented either based on geographical regions or based on boundaries between different areas. Regions with anatomical or functional characteristics may be categorized based on intensity patterns surrounded by a group of adjacent pixels [20].

This study employs region-based segmentation, a technique that partitions the area of interest (ROI) into subgroups based on distinct patterns or textures. K-means clustering utilizes the local mean as a cluster prototype to partition different interpretations into k groups. To identify clusters within the data, this approach aggregates all the groups represented by the variable k. The algorithm determines the nearest data point by calculating the squared Euclidean distances. The approach categorizes all data points into one of k categories based on the provided characteristics [21].

The K-NN classifier has been used in several research investigations. Pattern recognition employs a multitude of methods to classify items. K-NN utilizes the nearest training samples to classify objects. K-NN acquires knowledge from occurrences. Calculation is deferred until categorization is finished when a locally approximated function is used. KNN is most suitable as a classification approach when there is little knowledge about the historical distribution of the data. KNN is a widely used technique for categorizing patterns. Various observers have found that KNN computing is very effective when used to diverse sorts of data [22].

SVM, developed by Vatnik, has piqued the curiosity of scholars worldwide. The data is first divided into two groups using a support vector machine classifier. Once the classifier has been trained using the training data, a test model is created. At times, multiclass categorization tasks occur. This task will need a substantial number of binary classifiers. Based on several research, SVM demonstrates superior performance compared to other existing classification approaches. Support vector machines are capable of doing image categorization. According to the study, SVMs achieve much greater accuracy compared to other traditional classifiers [23].

13.4 Result Analysis

The researchers have assessed "Dataset-160 and Data-255" from the Harvard clinical college of "architecture" [24]. They assessed "Dataset-160," which includes "Normal-20," "Abnormal-140," and "Dataset-255", which includes 35 normal and 220 abnormal "T2-weighted" MR256x256 axial aircraft encephalon images, over the course of the investigation. Eleven different syndromes were represented by the "Dataset-255" Irregular Encephalon Magnetic Resonance metaphors, which connect to the "Dataset-160" by connecting the seven different syndromes. Among the seven disorders included in "Dataset-160" are Huntington's syndrome, Alzheimer's infection, Alzheimer's illness, and examples of agnosia, glioma, meningioma, Pick's infection, and sarcoma. "Dataset-255" contains images of four new syndromes: chronic subdural hematoma, herpes encephalitis, and a handful of other types of sclerosis. "Dataset-255" as shown in Figure 13.2: Result Comparison of Classifiers.

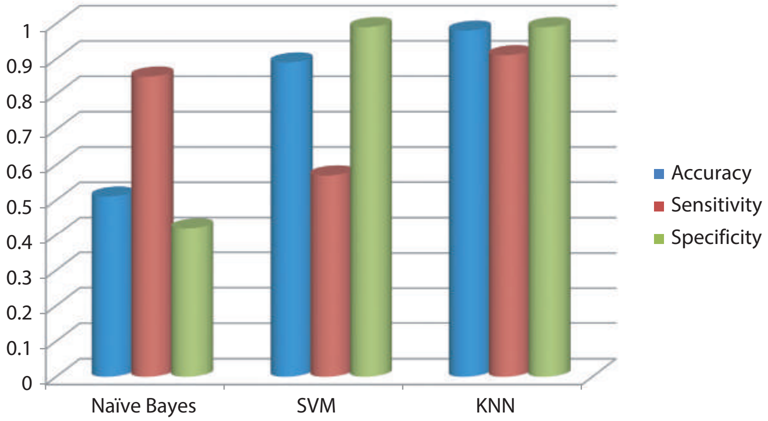


Figure 13.2 Result comparison of classifiers.

13.5 Conclusion

Across all age groups, the death rate has risen due to an increase in encephaloma tumors. Physical tumor identification is difficult and time-consuming for clinicians due to tumor complexity and the involution of noise in MR imaging data. Therefore, it is crucial to diagnose and localize the tumor early on. By combining medical images with segmentation and relegation methods, early diagnosis of malignant tumors at multiple stages becomes possible. To detect brain tumors, our method segments and categorizes MRI images using machine learning. Feature extraction, classification, segmentation, and support vector machine (SVM) and k-nearest neighbor (KNN) techniques are all part of this system. It is doing well, if KNN's accuracy is to be believed. When it comes to specificity, SVM and KNN are neck and neck. Compared to other algorithms, KNN performs better in terms of sensitivity.

References

1. Mohan, G. and Subashin, M., MRI based medical image analysis: Survey on brain tumor grade classification. *Biomed. Signal Process. Control*, 39, 139–161, 2018.
2. Işın, A., Direkoğlu, C., Şah, M., Review of MRI-based brain tumor image segmentation using deep learning methods. *Procedia Comput. Sci.*, 102, 317–324, 2016.
3. Menze, B.J., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R. *et al.*, The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging*, 34, 1993–2024, 2015.
4. Litjens, G., Kooi, T., Bejnordi, B.E., Arindra, A., Setio, A., Ciompi, F., Ghafoorian, M., van der Laak, J.A.W.M., van Ginneken, B., Sánchez, C.I., A survey on deep learning in medical image analysis. *Med. Image Anal.*, 42, 60–88, 2017.
5. Cheng, J., Huang, W., Cao, S., Yang, R., Yang, W., Yun, Z., Wang, Z., Feng, Q., Enhanced performance of brain tumor classification via tumor region augmentation and partition. *PLoS One*, 10, e0140381, 2015.
6. Sachdeva, J., Kumar, V., Gupta, I., Khandelwal, N., Ahuja, C.K., A package-SFERCB-Segmentation, features extraction, reduction and classification analysis by both SVM and ANN for brain tumors. *Appl. Soft Comput.*, 47, 151–167, 2016.
7. Suryavamsi, R.V., Sai Thejaswin Reddy, L., Saladi, S., Karuna, Y., Comparative Analysis of Various Enhancement Methods for Astrocytoma MRI Images.

International Conference on Communication and Signal Processing (ICCSP), pp. 0812–0816, 2018.

8. De Blasi, B., Galazzo, I.B., Pasetto, L., Storti, S.F., Koepp, M., Barnes, A., Menczaz, G., Pipeline Comparison for the PreProcessing of Resting-State Data in Epilepsy. *26th European Signal Processing Conference (EUSIPCO)*, pp. 1137–1141, 2018.
9. Poornachandra, S. and Naveena, C., Pre-processing of MR Images for Efficient Quantitative Image Analysis Using Deep Learning Techniques. *International Conference on Recent Advances in Electronics and Communication Technology (ICRAECT)*, pp. 191–195, 2017.
10. Widyarto, S., Kassim, S.R.B., Sari, W.K., 2Dsigmoid enhancement prior to segment MRI Glioma tumor: Pre image-processing. *4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, pp. 1–5, 2017.
11. Jui, S.-L., Zhang, S., Xiong, W., Yu, F., Fu, M., Wang, D., Hassanien, A.E., Xiao, K., Brain MRI Tumor Segmentation with 3D Intracranial Structure Deformation Features. *IEEE Intell. Syst.*, 31, 2, 66–76, 2016.
12. Zhang, J., Liu, M., An, L., Gao, Y., Shen, D., Alzheimer's Disease Diagnosis Using Landmark-Based Features From Longitudinal Structural MR Images. *IEEE J. Biomed. Health. Inf.*, 21, 6, 1607–1616, 2017.
13. Piantadosi, G., Marrone, S., Fusco, R., Sansone, M., Sansone, C., Comprehensive computer-aided diagnosis for breast T1-weighted DCE-MRI through quantitative dynamical features and spatio-temporal local binary patterns. *IET Comput. Vision*, 12, 7, 1007–1017, 2018.
14. Osadebey, M., Pedersen, M., Arnold, D., Wendel-Mitoraj, K., No-reference quality measure in brain MRI images using binary operations, texture and set analysis. *IET Image Proc.*, 11, 9, 2017.
15. Tsai, H.-Y., Zhang, H., Hung, C.-L., Min, G., GPU-Accelerated Features Extraction From Magnetic Resonance Images. *IEEE Access*, 5, 22634–22646, 2017.
16. Pereira, S., Pinto, A., Alves, V., Silva, C.A., Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images. *IEEE Trans. Med. Imaging*, 35, 5, 1240–1251, 2016.
17. Kim, J., Lenglet, C., Duchin, Y., Sapiro, G., Harel, N., Semiautomatic Segmentation of Brain Subcortical Structures From High-Field MRI. *IEEE J. Biomed. Health. Inf.*, 18, 5, 1678–1695, 2014.
18. Makropoulos, A., Gousias, I.S., Ledig, C., Aljabar, P., Serag, A., Hajnal, J.V., David Edwards, A., Counsell, S.J., Rueckert, D., Automatic Whole Brain MRI Segmentation of the Developing Neonatal Brain. *IEEE Trans. Med. Imaging*, 33, 9, 1818–1831, 2014.
19. Iqbal, Z., Khan, M., Sharif, M., Shah, J., ur Rehman, M., Javed, K., An automated detection and classification of citrus plant diseases using image processing techniques: A review. *Comput. Electron. Agric.*, *153*, 12–32, 2018, doi: 10.1016/j.compag.2018.07.032.

20. Zhao, W. and Wang, J., A new method of the forest dynamic inspection color image sharpening process. *2010 3Rd International Conference On Advanced Computer Theory And Engineering(ICACTE)*, 2010, doi: 10.1109/icacte.2010.5579715.
21. Reza, M., Na, I., Baek, S., Lee, K., Rice yield estimation based on K-means clustering with graph-cut segmentation using low-altitude UAV images. *Biosyst. Eng.*, *177*, 109–121, 2019, doi: 10.1016/j.biosystemseng.2018.09.014.
22. Deng, Z., Zhu, X., Cheng, D., Zong, M., Zhang, S., Efficient k NN classification algorithm for big data. *Neurocomputing*, *195*, 143–148, 2016, doi: 10.1016/j.neucom.2015.08.112.
23. Suryawati, E., Pardede, H., Zilvan, V., Ramdan, A., Krisnandi, D., Heryana, A. *et al.*, Unsupervised feature learning-based encoder and adversarial networks. *J. Big Data*, *8*, 1, 2021, doi: 10.1186/s40537-021-00508-9.
24. http://med.harvard.edu/AANLIB/
25. Nahar, S., Pithawa, D., Bhardwaj, V., Rawat, R., Rawat, A., Pachlasiya, K., Quantum Technology for Military Applications, in: *Quantum Computing in Cybersecurity*, pp. 313–334, 2023.
26. Pithawa, D., Nahar, S., Bhardwaj, V., Rawat, R., Dronawat, R., Rawat, A., Quantum Computing Technological Design Along with Its Dark Side, in: *Quantum Computing in Cybersecurity*, pp. 295–312, 2023.
27. Chauhan, D., Singh, C., Rawat, R., Chouhan, M., Conversational AI Applications in Ed-Tech Industry: An Analysis of Its Impact and Potential in Education, in: *Conversational Artificial Intelligence*, pp. 411–433, 2024.
28. Chauhan, D., Singh, C., Rawat, R., Dhawan, M., Evaluating the Performance of Conversational AI Tools: A Comparative Analysis, in: *Conversational Artificial Intelligence*, pp. 385–409, 2024.
29. Shrivastava, M., Patil, R., Bhardwaj, V., Rawat, R., Telang, S., Rawat, A., Quantum Computing and Security Aspects of Attention-Based Visual Question Answering with Long Short-Term Memory, in: *Quantum Computing in Cybersecurity*, pp. 395–412, 2023.
30. Rawat, R., Chakrawarti, R.K., Raj, A.S.A., Mani, G., Chidambarathanu, K., Bhardwaj, R., Association rule learning for threat analysis using traffic analysis and packet filtering approach. *Int. J. Inf. Technol.*, 15, 6, 3245–3255, 2023.

14

Optimized Machine Learning Techniques for Software Fault Prediction

Chetan Shelke[1]*, Ashwini Mandale (Jadhav)[2], Shaik Anjimoon[3], Asha V.[4], Ginni Nijhawan[5] and Joshuva Arockia Dhanraj[6]

[1]Alliance College of Engineering and Design, Alliance University, Bangalore, India
[2]Computer Science and Engineering, Rajarambapu Institute of Technology, Rajaramnagar, Shivaji University, Kolhapur, Maharashtra, India
[3]Institute of Aeronautical Engineering, Dundigal, Hyderabad, India
[4]Department of Computer Applications, New Horizon College of Engineering, Bangalore, India
[5]Lovely Professional University, Phagwara, India
[6]Department of Computer Science and Engineering (AI&ML), School of Engineering, Dayananda Sagar University, Devarakaggalahalli, Harohalli, Kanakapura Road, Ramanagara District, Bengaluru, Karnataka, India

Abstract

Software fault prediction serves to improve testing efficiency and software quality by enabling the early discovery of software problems. This is accomplished via improved program quality. In most cases, the procedure of categorizing is used for the purpose of error detection. During the classification process, coding features and other characteristics are used to produce predictions about the potential occurrence of mistakes. Due to the fact that software defect detection is heavily influenced by poor categorization judgments, it is necessary to have an improved decision-making model to anticipate trends by making use of the attributes retrieved from datasets. Through the use of a collection of software failure datasets, these researchers evaluate the performance of the offered techniques in comparison to that of a wide variety of machine learning classifiers. Accuracy, balance, area under the curve, false alarm rate, detection rate, or recall rate were some of the performance criteria that were used in the evaluation of the recommended approaches.

**Corresponding author*: Chetan.shelke@alliance.edu.in

Rajesh Kumar Chakrawarti, Ranjana Sikarwar, Sanjaya Kumar Sarangi, Samson Arun Raj Albert Raj, Shweta Gupta, Krishnan Sakthidasan Sankaran and Romil Rawat (eds.) Natural Language Processing for Software Engineering, (207–220) © 2025 Scrivener Publishing LLC

***Keywords*:** Scientific programming, machine learning, software fault prediction, software quality improvement

14.1 Introduction

Software faults are errors that prevent the software from functioning correctly and are referred to by that same term. Consequently, the fundamental purpose of the team should be to develop software that is free of mistakes whenever possible. Ineffective software that is both poorly written and poorly performed is a definite method to waste resources (including time, money, and energy). Even if there is no such thing as a perfect system, it is essential that there are as few errors as possible and that their negative effects are minimized. This will allow the system to be utilized more frequently and with a greater degree of assurance. When it comes to software modules, outputs that are wrong could be the result of poor-quality software as well as incorrect programming logic. It is possible that this will result in the client becoming dissatisfied, which will in turn push up the cost of maintenance. A decrease in customer satisfaction may drive the client to urge that the developer solve the issue, or the client may opt to withdraw from the contract entirely. Both of these outcomes are possible. In the end, poor quality and the failure of the project are both potential outcomes that can result from defects. When it comes to software development, metrics not only make it possible to evaluate the effectiveness and quality of a project, but also make it easier to strategically distribute available resources [1].

As a result of the fact that it is not always possible to discover and correct faults while the development process is in progress, it is essential to keep an eye out for them and correct them as soon as you or your teammates see them. It is possible for errors to occur at any stage of the software development process because it is an iterative process. Therefore, to produce a maintenance strategy, it is essential to discover problems in software modules in a methodical manner prior to the deployment of the project. Software development is a process that has the potential to produce a variety of different types of errors. It is possible that it is foolish to believe that these vulnerabilities are introduced into the system at the beginning of the cycle and that they can be addressed as the cycle progresses. Every single stage of the process is open to the chance of making a mistake. The term "software defects" refers to irregularities that occur throughout the process of developing software and result in the software not conforming to the standards that were anticipated [2].

For project managers, the goal of accurately predicting software development cycle (SDLC) errors in a timely way is an ideal goal, but it is also a very tough goal to achieve [3]. In this day and age, it is quite challenging to produce software that is not only dependable but also free of faults, particularly during the development period. During the process of development, there are a number of problems that get increasingly intricate and ambiguous. It is an extremely challenging endeavor to either avoid or deal with all of these problems. It is impossible to guarantee that there will be no errors in the process, regardless of how much preparation or documentation is done. The product's quality suffers as a result of these defects, and customers are dissatisfied with the purchase. Classification and prediction are two independent concerns that can be utilized to express software fault prediction: classification and prediction. The first method can be utilized to assign defect labels to data classes, while the second method can be utilized to create predictions regarding trends on a continuous scale. The distinction between the two kinds of predictions is that one approach makes use of discrete, unordered labels or categorical variables, while the other approach makes use of functions that have continuous values. Through the utilization of this kind of analysis, software defects may be comprehended and treated in a more effective manner [4].

Software defect prediction is a learning issue that involves predicting the chance of software modules having flaws based on previous data. This approach is known as software defect prediction. To derive the defect prediction models, the log files of older versions are analyzed to extract the qualities that correspond to the code. Only by making use of predictions is it possible to arrive at an exact count of the number of defects that are present in the tested module as well as the locations of those defects. Sadly, this requirement has not been satisfied because of the inconsistencies that have been recorded in the defect recording procedure as well as the very complicated nature of the prediction process [5].

Most of the time, SDP models are able to determine whether or not a specific collection of test data contains any errors. Metrics for software can be utilized to ascertain the degree to which the software is prone to defects. In software development process (SDP) models, the state of the software bug is used as the dependent variable. These models make use of independent variables that are acquired and evaluated throughout the software development lifecycle. In the context of software, metrics refer to measurements that are taken of either the software itself or the specifications of the software. Utilizing this is helpful when attempting to quantify the quality of the curriculum. "Continuously applying techniques based on measurement in the process of developing software and its products and in providing meaningful and timely Management Information (MI) along with the methods to improve its products and processes" is the resultant definition of metrics for software development [6].

In numerous occasions, software quality prediction models are constructed using software metrics and defect data from a software project that has already been developed and deployed. This makes it possible to evaluate the development module's fault proneness, which is a significant benefit. A non-defect forecast implies that the current module is of good quality, which provides the opportunity for the resources to be pooled for some other activity. On the other hand, a defect prediction that results in the requirement to shift resources to address the issue is the reverse of what is intended. Our ultimate objective is to find a balance between quality and dependability while making the most of the resources we have currently available to us. As a result of the fact that the model is dependent on the information that is contained in the underlying software measures, selecting the appropriate collection of software measures is an essential component of the process of building a model for computing software quality prediction. Due to the fact that software companies are subject to a penalty when customers report software defects, it is beneficial to have knowledge of the expected number of errors that must be present in a software product at a particular stage of development. To determine when it is appropriate to release a software product, it is necessary to have the ability to forecast the number of issues that will arise at any given stage [7].

Machine learning techniques that are now in use include classification and clustering, which are two of the most prevalent methods. Having characteristics or data of a good quality is absolutely necessary for the operation of data models. If the data is faulty or contains information that is not relevant to the problem at hand, then the accuracy of the model's predictions will be undermined. As a consequence of this, it is of the utmost importance to prepare the data in such a way that it can be less redundant and free of noise. The procedure of selecting features is quite important. Classification is a well-liked method for predicting software defects. It employs a model that is constructed using metrics data of software that has been used in the past to classify the codes of software attributes according to whether or not they are defective. On account of the fact that the classifier is typically the weakest link due to the limited number of features that are used for prediction, it plays an extremely important role in the accurate detection of defects [8].

14.2 Literature Survey

It was recommended that a three-way decision structure be used for the purpose of anticipating software flaws. This decision-making framework takes into consideration costs, and it does so by using the well-known two-stage categorization paradigm as its base. A system that utilizes attribute selection for defect prediction was developed by the authors. This system is built on five different classifiers: Random Forest, Random Tree, LWL, IBk, and KStar. In this study, we evaluated two separate classification learners using six alternative feature selection procedures. Additionally, we assessed approaches that coupled feature selection with data sampling, as well as cases in which feature selection was used on its own. It has been shown via the results of the experiments that the SelectRUSBoost technique greatly increases classification performance when compared to other strategies that have been used in the past. When comparing the performance, the ROC values and accuracy are taken into consideration. According to the findings, the framework was able to successfully cut down on the number of characteristics that were used for each dataset by an average of six times. In addition, it was discovered that LWL performed better than the other four classifiers when it was put through a test with a 10CV or a 10 Cross-Validation and a split percentage of 66%. A validation accuracy grade that is exceedingly low suggests that the model is unable to function properly when dealing with complex data [9].

A Twin Support Vector Machine (TSVM) was used by the writers to make a prediction about the mistakes that would be present in a software product that has been updated. This prototype was far more effective than the ones that came before it. Additionally, when compared to all of the SDP approaches that were offered, the TSVM with a Gaussian kernel function has shown the highest level of efficiency. To maintain a high level of quality in the program, the author was also able to anticipate the shortcomings of the new technique. During the testing phase, the strategy that was advised was already more cost-effective. On the other hand, it was shown that the model of huge data displays very low levels of scalability [10].

The authors proposed the use of a multilayer neural network as a means of significantly enhancing the capacity of learning algorithms to forecast software issues. This article presents a unique technique that blends evolutionary algorithms and the concepts of support vector machines (SVM). With the help of this innovative method, the classification margin was increased to its maximum, and overfitting was avoided. Using three different datasets from NASA, eleven different machine learning prototypes and statistical methodologies were put through their paces to establish how effective the aforementioned procedure was. It was concluded that the process described above offered a higher level of trust and authenticity when compared to the other prototypes. In spite of this, it was found that neural networks greatly increased the degrees of complexity that the models had [11].

For the purpose of determining FCRelevance, the authors used Chi-Square, Relief, and Information Gain. They also quantified FFCorrelation by using Symmetric Uncertainty. To analyze the rate of redundancy and, as a last step, to evaluate the effectiveness of defect predictors, empirical tests were carried out. As a basis for putting the proposed framework into action, real-world projects were used such as NASA and Eclipse. This endeavor resulted in a final product that proved the effectiveness of the framework that was provided and gave advice for achieving feature selection that is cost-effective. On the other hand, the model could not provide satisfactory results when the levels of uncertainty measure were low or equivalent [12].

TSCS is a two-stage cost-sensitive learning technique that combines cost data before, during, and after the feature selection and classification processes. The authors suggested this approach for the SDP. Following that, the feature selection process resulted in the development of three cutting-edge algorithms that were very sensitive to cost. By including cost information into the traditional feature selection methods, it was feasible to generate the Cost-Sensitive Variance Score (CSVS), the Cost-Sensitive Laplacian Score (CSLS), and the Cost-Sensitive Constraint Score (CSCS). These scores could be calculated. There is an increase in the computational complexity when more than one model is used. An additional review of the approaches that have been presented is carried out with the use of seven real data sets that have been generated from NASA investigations. According to the results of the experiments, TSCS performed much better than other cost-sensitive methods when it came to forecasting software problems. During the feature selection process, it was found out that the recommended ways performed better than the other standard methods. This was done to further assure that the utilization of cost information would be effective [13].

A concept for active learning has been developed as a result of research into automating the models that academics have developed to improve their ability to predict which releases would include flaws. When compared to its equivalent learning technique, the results demonstrated that active learning integration with uncertainty sampling performed much better. When it came to discovering and integrating the relevant information in the datasets that were supplied, the scientists discovered that multidimensional scaling with random forest similarity was more successful than feature selection. Improvements were made to the method as a result of three successive versions of Eclipse's flawed module prediction [14].

It was stated that a mechanism existed for forecasting software problems by making use of certain metrics. The authors gave feature selection strategies a great deal of attention and proposed a novel strategy that took a hybrid approach to the problem. The findings indicated that the strategy that was recommended improved the accuracy of categorization. A study was conducted to investigate the influence that feature selection techniques, metrics sets, and dataset size had on the prediction of software defects. One of the drawbacks of this technology is that it allows for the use of several models, which leads to very high levels of computational complexity. The research that was done on the subject demonstrated that support vector machines (SVM) and other soft computing methods provide superior classification accuracy when compared to other classifiers. On the other hand, the performance of SVM is NP-Hard and is reliant on the parameter selection of the kernel module. Despite the fact that grid-based methods and meta heuristic approaches, such as Genetic Algorithm (GA), were successfully used, the solutions that were obtained were not ideal. Additionally, GA had a slow convergence rate [15].

The authors presented a strategy that comprised methods for data sampling that were aimed at addressing class imbalance, as well as a method for feature selection that was focused on picking significant characteristics. To examine software engineering, especially the use of classification models for the purpose of software quality prediction, this investigation was carried out. Whenever data sampling and feature selection are combined, there are a number of considerations that need to be taken into account. This case study makes use of nine different software measurement datasets that were taken from the repository of the PROMISE project. The empirical findings revealed that feature selection making use of sampled data worked better than feature selection making use of original data, and that models for defect prediction functioned equally regardless of the data that was used for training. Although the selection of features is significant, it is not sufficient to determine the accuracy of predictions; the prediction model is of more significance on this front [16].

14.3 Methods

Machine learning procedure for software fault prediction is shown below in Figure 14.1.

It is a usual practice to combine the firefly approach with the ant colony optimization (ACO) algorithm when feature selection operations need to be carried out. This is because the firefly technique is so effective. When it comes to intrusion-related data, the firefly approach is responsible for selecting sub-bands, while the ACO strategy is in charge of managing classification and feature continuity analysis. With the help of the ACO approach, both of the objectives are achieved. One example of such a comparison is the approach that algorithm developers use, which is analogous to the way in which ants construct and upgrade their nest spaces [17].

The support vector machine (SVM) is the most appropriate classifier among others many others since it can be utilized in a wide variety of contexts, including medical diagnosis and identification, pattern recognition, text classification, organism identification, and Chinese character classification, among others and many more. Through the construction of an N-dimensional hyperplane, support vector machines (SVMs) divide the data into two distinct groups: those with the most favorable outcomes are located on one side of the hyperplane, while those with less favorable outcomes are located on the other side. By maximizing the margins between

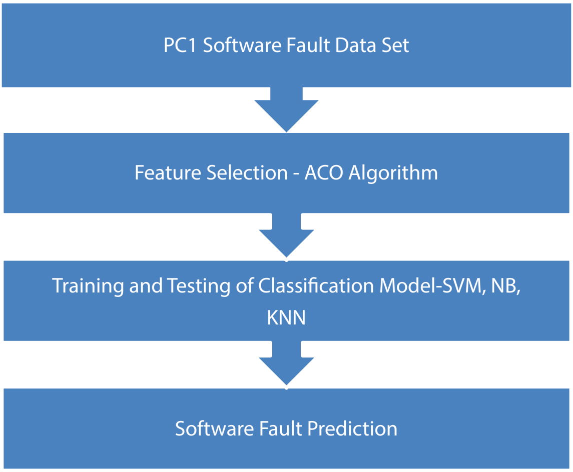

PC1 Software Fault Data Set
Feature Selection - ACO Algorithm
Training and Testing of Classification Model-SVM, NB, KNN
Software Fault Prediction

Figure 14.1 Machine learning for software fault prediction.

support vectors and managing non-linear regions, support vector machines (SVM) map data into a space that is not uniform. The hyperplane then executes separation as a result of this mapping [18].

KNN [19] is the most straightforward and effective method for data categorization that everyone can utilize. It is also the most straightforward. In addition to being easy to use and delivering results that are highly competitive, it also has the capability of implicitly computing decision boundaries. Considering that this approach is dependent on having a selection of instances to pick from, there is a possibility that disputes over which characteristics are relevant and which are not would result in findings that are less accurate. To establish whether or not a dataset is n-dimensional, the authors first sum up all of the samples that are included inside the dataset, and then we examine the vector that is produced to determine whether or not it is n-dimensional.

Bayesian classifiers are a kind of statistical classifier [22–24] that makes predictions about the appropriate classes to which data points should belong by assessing the likelihood of class membership. Bayesian classifiers are used in the application of statistics. In a manner that is analogous, the Naive Bayes [25–27] classification approach is significantly dependent on optimum rules; this theorem serves as the basis for the method. There are a number of clear benefits associated with doing so, including the fact that it is easy to put into action and brings about outcomes that can be relied upon. In this study, the authors want to investigate and compare the classification methods used by a number of different convolutional neural network classifiers, as well as the decision tree classifier and the naive Bayes classifier classification system. Under the assumption that the prior probabilities P(c), P(x), and P(x | c) are known, the Bayes theorem may be used to determine the posterior probability P (c | x). Due to the fact that it is predicated on the premise that the effect of a given predictor value (x) on a certain class (c) stays constant across all potential values, the Naive Bayes classifier continues to carry out its typical operations. The concept of "conditional independence" immediately comes to mind when attempting to provide an explanation for this set of circumstances. Among the essential actions involved in the production of a frequency table are the determination of the posterior probability and the classification of characteristics according to the frequency with which they occur. After that, the Naive Bayesian equation is used to get the posterior probability for each of these classes. Those students who belong to the category for whom this prognosis has the greatest likelihood of being right will be the ones to profit the most from its achievement [20]. The Figure 14.2 shows about the Result comparison.

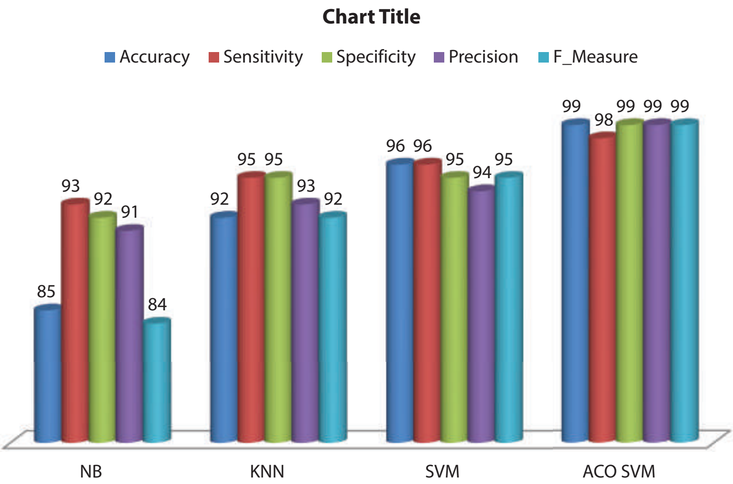


Figure 14.2 Result comparison.

14.4 Result Analysis

The PC1 data collection is used in the scientific research that is being carried out [21]. There are data errors that belong to a specific NASA program that are included in this collection of data. This particular dataset comprises 1,192 records. The data set from PC1 is used as the input for this design. Ant colony optimization is a strategy that selects the characteristics that are most significant. During the training process, the support vector machine is trained using the features that have been chosen in advance. For the purposes of SVM training and testing, the PC1 dataset is used. The performance of the ACO SVM is examined in conjunction with the SVM, Naive Bayes, and KNN classifiers in order to establish its effectiveness.

14.5 Conclusion

In the course of the study, a number of different aspects were investigated, including issues pertaining to the quality of the data, metrics for performance evaluation, metrics for software, and failure prediction. The findings of this study shed light on a range of methodological challenges that were related with software fault prediction jobs. The usage of feature extraction for data with a high dimensionality and classification for data with a data class imbalance associated to software quality issues has been widespread. Both of these techniques have been used extensively. There are a few different metrics that are used to examine the outcomes to determine how accurate the predictions are. The study also revealed difficulties that researchers may examine further to enhance the software defect prediction approach. This was emphasized in order to improve the system.

References

1. Mishra, A., Shatnawi, R., Catal, C., Akbulut, A., Techniques for Calculating Software Product Metrics Threshold Values: A Systematic Mapping Study. *Appl. Sci.*, 11, 11377, 2021.
2. Özakıncı, R. and Tarhan, A., Early software defect prediction: A systematic map and review. *J. Syst. Software*, 144, 216–239, 2018.
3. Son, L.H., Pritam, N., Khari, M., Kumar, R., Phuong, P.T.M., Thong, P.H., Empirical Study of Software Defect Prediction: A Systematic Mapping. *Symmetry*, 11, 212, 2019.
4. Bhavana, K., Nekkanti, V., Jayapandian, N., Internet of things enabled device fault prediction system using machine learning, in: *International Conference on Inventive Computation Technologies*, pp. 920–927, Springer, Cham, Switzerland, 2019.
5. Ramakrishnan, R. and Kaur, A., An empirical comparison of predictive models for web page performance. *Inf. Software Technol.*, 123, 106307, 2020.
6. Padhy, N., Satapathy, S.C., Mohanty, J., Panigrahi, R., Software reusability metrics prediction by using evolutionary algorithms: The interactive mobile learning application RozGaar. *Int. J. Knowledge-Based Intell. Eng. Syst.*, 22, 261–276, 2018.
7. Wang, P., Jin, C., Jin, S.W., Software defect prediction scheme based on feature selection. *International Symposium on Information Science and Engineering (ISISE)*, IEEE, pp. 477–480, 2012, DOI: 10.1109/ISISE.2012.114.
8. Reena, P. and Rajan, B., A Novel Feature Subset Selection Algorithm for Software Defect Prediction. *Int. J. Comput. Appl.*, 100, 17, 39–43, 2014.
9. Jing, X.Y., Ying, S., Zhang, Z.W., Wu, S.S., Liu, J., Dictionary learning based software defect prediction. *Proceedings of the 36th International Conference on Software Engineering*, ACM, pp. 414–423, 2014, DOI:10.1145 / 2568225.2568320.
10. Xu, Z., Xuan, J., Liu, J., Cui, X., MICHAC: Defect Prediction via Feature Selection based on Maximal Information Coefficient with Hierarchical Agglomerative Clustering. *IEEE 23rd International Conference on Software Analysis, Evolution, and Reengineering (SANER)*, vol. 1, IEEE, pp. 370–381, 2016.

11. Xia, Y., Yan, G., Jiang, X., Yang, Y., A new metrics selection method for software defect prediction. *International Conference on Progress in Informatics and Computing (PIC)*, IEEE, pp. 433–436, 2014, DOI:10.1109/PIC. 2014 .6972372.
12. Mandal, P. and Ami, A.S., Selecting best attributes for software defect prediction. *IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE)*, IEEE, pp. 110–113, 2015, DOI: 10.1109/ wiecon-ece.2015.7444011.
13. Dejaeger, K., Verbraken, T., Baesens, B., Toward comprehensible software fault prediction models using bayesian network classifiers. *IEEE Trans. Software Eng.*, 39, 2, 237–257, 2013.
14. Bishnu, P.S. and Bhattacherjee, V., Software fault prediction using quad tree-based k-means clustering algorithm. *IEEE Trans. Knowl. Data Eng.*, 24, 6, 1146–1150, 2012.
15. Gondra, I., Applying machine learning to software fault-proneness prediction. *J. Syst. Software*, 81, 2, 186–195, 2008.
16. Catal, C., Software fault prediction: A literature review and current trends. *Expert Syst. Appl.*, 38, 4, 4626–4636, 2011.
17. Malik, V., Mittal, R., Singh, J., Rattan, V., Mittal, A., Feature Selection Optimization using ACO to Improve the Classification Performance of Web Log Data. *2021 8th International Conference on Signal Processing and Integrated Networks (SPIN)*, pp. 671–675, 2021, doi: 10.1109/ SPIN52536.2021.9566126.
18. Shafaghat, M. and Dezvareh, R., Support vector machine for classification and regression of coastal sediment transport. *Arabian J. Geosci.*, 14, 2009, 2021, https://doi.org/10.1007/s12517-021-08360-0.
19. Goyal, S., Handling Class-Imbalance with KNN (Neighbourhood) Under-Sampling for Software Defect Prediction. *Artif. Intell. Rev.*, 55, 3, 2023–2064, 2021, Available: 10.1007/s10462-021-10044-w.
20. Rahim, A., Hayat, Z., Abbas, M., Rahim, A., Rahim, M.A., Software Defect Prediction with Naïve Bayes Classifier. *2021 International Bhurban Conference on Applied Sciences and Technologies (IBCAST)*, pp. 293–297, 2021, doi: 10.1109/IBCAST51254.2021.9393250.
21. https://www.openml.org/search?type=data&sort=runs&id=1068&status=active
22. Pithawa, D., Nahar, S., Bhardwaj, V., Rawat, R., Dronawat, R., Rawat, A., Quantum Computing Technological Design Along with Its Dark Side, in: *Quantum Computing in Cybersecurity*, pp. 295–312, 2023.
23. Shrivastava, M., Patil, R., Bhardwaj, V., Rawat, R., Telang, S., Rawat, A., Quantum Computing and Security Aspects of Attention-Based Visual Question Answering with Long Short-Term Memory, in: *Quantum Computing in Cybersecurity*, pp. 395–412, 2023.
24. Rawat, R., Chakrawarti, R.K., Raj, A.S.A., Mani, G., Chidambarathanu, K., Bhardwaj, R., Association rule learning for threat analysis using traffic

analysis and packet filtering approach. *Int. J. Inf. Technol.*, *15*, 6, 3245–3255, 2023.

25. Rathi, M. and Rajavat, A., High Dimensional Data Processing in Privacy Preserving Data Mining, in: *2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT)*, 2020, April, IEEE, pp. 212–217.
26. Patsariya, M. and Rajavat, A., Network Path Capability Identification and Performance analysis of Mobile Ad hoc Network, in: *2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT)*, 2020, April, IEEE, pp. 82–87.
27. Dhar, S., Dhar, U., Rajavat, A., Factors and Attributes of Team Players: A Study of Engineering Students in India, in: *International Simulation and Gaming Association Conference*, pp. 53–60, Springer International Publishing, Cham, 2021, September.

15

Pancreatic Cancer Detection Using Machine Learning and Image Processing

Shashidhar Sonnad[1]*, Rejwan Bin Sulaiman[2], Amer Kareem[3], S. Shalini[4], D. Kishore[5] and Jayasankar Narayanan[6]

[1]Electronics and Communication Engineering, Sharnbasva University, Kalaburagi, Karnataka, India
[2]Department of Computer and Information Sciences, Northumbria University, Newcastle, UK
[3]Department of Computing, University of Wales Trinity Saint David, Swansea, UK
[4]Department of Mathematics, J.N.N Institute of Engineering, Ushaa Garden, Kannigaipair, Chennai, India
[5]Department of ECE, Aditya College of Engineering & Technology, Surampalem, India
[6]Department of Pharmacology, SRM College of Pharmacy, SRM Institute of Science and Technology Kattankulathur, Chengalpattu District, Tamil Nadu, India

Abstract

The absence of uniformity in data sources poses challenges for clinical diagnosis. The healthcare industry developments have led to enhanced diagnostic benefits. Medical professionals are optimistic that radiologists will enhance their ability to analyze images and provide accurate diagnoses through the utilization of computer-aided diagnosis technologies. Humans have always been susceptible to many potentially lethal illnesses. Pancreatic adenocarcinoma (PDAC) is a particularly prevalent and lethal illness among those aged 45 and older. PDAC is generally ranked as the fourth most common type of malignancy. CT, MRI, PET scans, and ultrasounds are the often used diagnostic methods for diagnosing disorders. To predict individuals' risk levels and detect symptoms of pancreatic disease for timely surgical preparation, this enables the creation of a reliable diagnostic system for pancreatic cancer using diverse clinical data, such as Structured EHRs, images of jaundiced eyes, and pancreatic CT scans. Extracting patterns from multi-modal

**Corresponding author*: shashidharsonnad1@gmail.com
Jayasankar Narayanan: ORCID: https://orcid.org/0000-0002-6149-2761

Rajesh Kumar Chakrawarti, Ranjana Sikarwar, Sanjaya Kumar Sarangi, Samson Arun Raj Albert Raj, Shweta Gupta, Krishnan Sakthidasan Sankaran and Romil Rawat (eds.) Natural Language Processing for Software Engineering, (221–232) © 2025 Scrivener Publishing LLC

healthcare data is a challenging task due to the pancreas's complex form, volume, and its position in the abdomen. Additionally, there are difficulties in identifying the pancreas, dealing with noisy data, and managing many risk factors associated with the disease. Extracting important disease patterns is currently extremely tough due to the existing state of affairs. Detecting tumors in pancreatic CT images and jaundice eye photographs is a difficult task with the existing segmentation methods. This study aims to extract meaningful disease patterns from clinical data to diagnose pancreatic ailment, taking into account the variety of medical data sources and the lack of a common framework.

***Keywords*:** Pancreatic cancer, machine learning, classification, detection, PSO SVM, accuracy

15.1 Introduction

Pancreatic cancer is one of the leading causes of death from cancer all over the world. In patients who are above the age of 50 and have just been diagnosed with diabetes, there is an increased likelihood that they will develop pancreatic adenocarcinoma (PDAC). About eighty-five percent of the cases were caused by adenocarcinoma of the pancreas, and jaundice is a condition that happens before this cancer develops. The development of pancreatic cancer is a consequence of a wide range of diseases that lead to the malfunctioning of the pancreas. A significant number of these diseases are brought on by structural, social, and environmental changes that occur over the course of time. There are a variety of disorders that can affect the pancreas, including acute pancreatitis, chronic pancreatitis, exocrine pancreatic insufficiency, diabetes, cystic fibrosis, and hereditary pancreatitis. On the other hand, the most significant risk factors for pancreatic cancer are the use of tobacco products, diabetes, exposure to chemicals, heredity, bad eating habits, and excessive intake of alcohol. Pancreatic adenocarcinoma is the fourth most common type of cancer that results in death as a result of cancer. Electronic health records (EHRs) contain information about medical imaging (including CT, MRI, and PET scans, among others) as well as multimodal health clinical data (relating to pancreatic cancer) [1].

A technique known as machine learning is utilized in the process of developing algorithms for disease diagnosis and identification of severity. Following the confirmation of the presence of pancreatic cancer by a patient's scleral bilirubin level, a computed tomography (CT) scan is performed. The stage of the cancer as well as its severity can be determined by CT scans. Medical CT scans in grayscale and color are the inputs for the system that is used for diagnosing pancreatic cancer. Following the extraction of the original data, biomarkers or regions of interest are identified. Considering the pancreatic form and its placement below the stomach, the process of separating the cancer from the normal tissue is a challenging one [2]. The study of computed tomography (CT) scans makes it possible to differentiate between cancerous and non-cancerous tissues, as well as to stage cancer to plan therapies and procedures. How well a diagnostic system is able to recognize patterns of sickness based on patient records can be evaluated using performance metrics such as sensitivity and specificity. In most cases, the metrics of sensitivity, specificity, and false positive rate (FPR) are considered to be adequate for assessing a computer-aided pancreatic cancer diagnostic system. Using multi-modal clinical data, such as structured electronic health records, pictures of jaundiced eyes, and pancreatic computed tomography scans, this study presents strategies for improving the system's performance in terms of sensitivity, specificity, F-score, and error rate in the diagnosis of pancreatic cancer. These strategies are presented in this study [3].

There are two types of health information that are included in clinical data: organized and unstructured health information, as well as medical images. Some of the most common imaging modalities that are used for the diagnosis of pancreatic cancer include scleral recognition, positron emission tomography (PET) scans, ultrasound, computed tomography (CT), magnetic resonance imaging (MRI), and computed tomography (CT). Computed tomography (CT) scans are a type of imaging that involve the use of a focused beam of X-rays that rotates around the body of the patient in a circular motion. It is possible to create cross-sectional views, also known as slices, of the human anatomy through the use of computer processing of the images of the internal organs of the body that were obtained during the scan. Tomographic images, as opposed to traditional X-rays, are given their name due to the fact that these slices produce information that is specific to the images. CT scans of the pancreas are noninvasive methods that can be used to assess a patient's likelihood of developing cancer. These scans involve the collection or stacking of several images of the organ to produce a three-dimensional image that can detect abnormalities such as tumors [4]. One image showing pancreatic cancer is shown in Figure 15.1.

Images of pancreatic cancer obtained using CT and MRI are free of noise caused by patient movement, miscalibration of the detector, and other forms of interference. The noise removal function is essential to the process of image processing since it is necessary for the analysis of the picture. The reduction of noise and the preservation of data can be accomplished by a number of different ways, including linear and non-linear denoising techniques. In addition to protecting the data contained within the image,

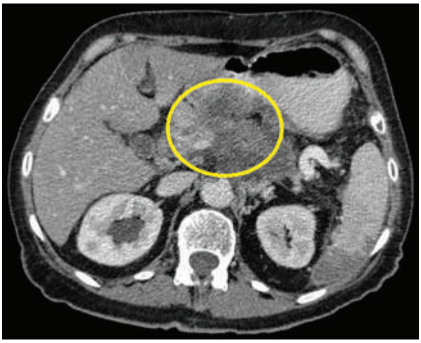


Figure 15.1 Pancreatic cancer tissue in CT scan image.

non-linear filters also give the edges more prominence [5]. Depending on the kind of noise that is present in the image, the proper filter is applied to remove the noise from the image. One can observe the progression of pancreatic tumors by looking at Figure 15.2.

The utilization of ensemble-based classifiers results in a significant improvement in the accuracy of pancreatic cancer classification systems. Learning strategies that are either supervised or unsupervised can be utilized to complete the task of classifying pancreatic segments simultaneously. Supervised learning is used to accomplish the classification of the regions of tumors. The system is intended to acquire knowledge based on the class that is defined for the image. It organizes the test data in accordance

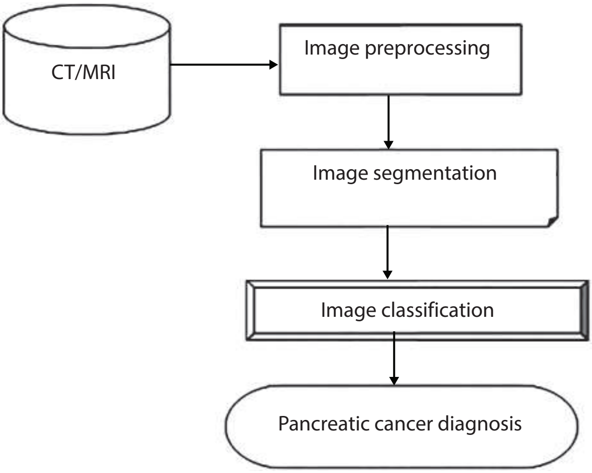


Figure 15.2 Pancreatic cancer detection process.

with the classification that was learned from the structure. During the categorization process, the region of the pancreas that contains the cancer is identified using picture analysis [6].

15.2 Literature Survey

There are two categories of pancreatic tumors: benign and malignant. These categories were determined by utilizing three models that were established [7]: LDA and KNN. Identifying patients who have malignant pancreatic tumors was the motivation for the development of the novel approach, which was developed with the intention of establishing the patient's existence. PLS is utilized in the process of removing Raman spectra data from pancreatic tumors.

Behrouz Alizadeh Savareh and colleagues [8] created a reliable diagnosis model for pancreatic cancer by utilizing miRNA biomarkers. This model was developed utilizing machine learning techniques. Using feature selection methods, the findings assign a score to the miRNAs, and they supply the index with miRNAs that have been chosen for their clinical value as biomarkers for the identification of pancreatic cancer. This demonstrated that the miRNAs with the highest degree of discrimination for the identification of pancreatic cancer have been obtained.

The research conducted by Abu Z. M. Dayem Ullah and colleagues [9] was one of the studies that investigated the link between pancreatic cancer and clinical factors throughout the course of time. It was not possible to conduct an accurate risk assessment for pancreatic cancer in connection to the emergence of co-morbidities and activities that are performed on a daily basis. The judgements may notify risk-stratification methods and improved investigation to persons who are at a higher risk. This is in addition to the fact that they provide systematic recognition of the target population for prospective cohort-based early detection research innovation.

Sk Md Mosaddek Hossain and colleagues [10] laid up the foundation for finding significant regulatory genes and important gene modules by using data from PDAC microarrays. Important regulators were able to examine the patterns and characteristics of gene activity connected to the progression of cancer as a result of this. In addition, the functionalities of the GAN should be rebuilt based on the expected top genes. The difference of partial correlation technique was utilized to accomplish the goal of identifying GAN based on gene function characteristics. Additionally, the recognition of the dynamical GENIE3 (dynGENIE3) technique was made possible by the use of gene regulatory inference.

An investigation was conducted by V. Jackins and colleagues [11] into the application of a Naive Bayes model in a number of disorders, including cancer, with the purpose of determining whether or not a patient was concerned about their illness. An investigation was carried out to analyze and contrast the disease data that were utilized in the model. In addition to demonstrating the efficiency of the classification algorithms, the outcomes of the simulations also highlight the intricacy of the datasets that were utilized.

A bagging of multiple Gaussian mixture approach was developed by Yaoyao Li and colleagues [12] to differentiate between samples that were normal and those that were related to tumors. For the purpose of slicing genomic sequences, it employs a coverage ratio as a rudimentary criterion. Following that, a Gaussian mixture approach was developed to determine the meaning of hazy copy number states. The bagging models were able to detect copy number gains and losses with a great deal greater precision as a result of improvements made to their sensitivity and specificity.

An AdaBoost ensemble method was developed by Sourabh Shastri and colleagues [13] as a part of their research on disease detection in classification. Both imputes and filters were employed by the erythemato-squamous diseases (ESD) team to ensure that the dataset remained in a balanced perspective. To complete the building, both the GB ensemble approach and a method that did not need it were utilized. A shared meta-classifier was utilized by the classes that were part of the GB ensemble technique to generate a variety of static base-classifier combinations possible. Stephen P. Pereira and colleagues [14] presented a review that was developed on the topic of early diagnosis of pancreatic cancer. They investigate the idea of making use of social media. Additionally, his work investigates the application of artificial intelligence in medical imaging and how it is connected to the goals of early discovery. Individualized pre-diagnostic tools for the discovery and validation of biological and epidemiological indicators can be obtained through the construction of new cohorts that are tailored to specific needs.

The identification of pancreatic tumors was spearheaded by Santosh Reddy P and Chandrasekar M when they were working from the ensemble technique [15]. Through the implementation of the intended strategy, the classifiers' capacity to identify pancreatic cancer is improved. The limited visual scheme of mammals serves as the foundation for the development of intelligent artificial picture detection classification. The study that was done by Md. Manjurul Ahsan and his colleagues [16] became the foundation for illness analysis performed by machines. In the beginning, the data about WOS were collected through the use of bibliometric research that was based on published works. There are a number of different domains that are encompassed by the most recent MLBDD algorithms. Some of these domains include kinds of diseases, data categories, applications, and estimating parameters.

15.3 Methodology

Despite the fact that the SVM technique is most frequently utilized for the purpose of achieving classification objectives, it is also capable of addressing challenges associated with regression. By locating a hyperplane, the data can be divided into a number of different groups. The algorithm attempts to maximize the distance between the class data points to locate the hyperplane that is going to be the most effective. This property of maximizing distance is referred to as margin maximization [17]. There are two types of support vector machine algorithms: the linear kind and the non-linear kind. The linear kind is the more common approach. A linear support vector machine classifies the training data by employing a hyperplane as the classification mechanism. In contrast, a non-linear support vector machine (SVM) is built by first substituting each dot product with a non-linear kernel function and then applying the kernel trick to maximum-margin hyper-planes. This process is repeated until the SVM [21–25] is complete. SVM performs exceptionally well in classification challenges that are nonlinear and complex. Furthermore, it provides a reliable form of extrapolation to the data that are uncertain. The current investigation makes use of support vector machines (SVMs) that are fitted with kernels referred to as "linear," "sigmoid," and "radial basis function (RBF)"[25–27] to enhance the categorization of histopathological images and to determine whether or not the characteristics that are retrieved are linearly discriminable or non-linear.

It is a graphical tool that employs branching methodology to represent all of the alternative choice outcomes that are depending on particular variables. This tool is known as a decision tree. Each node within a decision tree represents a feature; each branch represents a decision rule, and ultimately, each leaf node represents an outcome, also known as a target variable. The root node is the first node in a decision tree [24, 25, 28, 29]. It is possible to improve communication and make judgments in the face of uncertainty with the assistance of this method, which provides a visual depiction of a chosen condition. Another method which assists data scientists in arriving at sound findings is by providing them with the capability to perform computations in both the forward and backward directions.

In addition, decision trees are flexible enough to accommodate errors and data that are missing with relative simplicity. It is most appropriate for the method to use instances that are provided in the form of attribute-value pairs. In the context of decision trees, a classification tree or a regression tree are both valid options. A number of different industries, such as banking, finance, remote sensing, and healthcare, stand to gain from the implementation of the algorithm [18].

The Random Forest algorithm works by generating a forest of decision trees by using a randomly selected sample of data [19]. The bagging method is the most used training approach for the Random Forest algorithm. Using a random sample, the model is trained in an iterative manner, and the best prediction is derived by aggregating the results of all of the decision trees. The accuracy of our technique is not affected by outliers, and it continues to function normally even when there is no data available. One of its most notable strengths is that it can process numerical, category, and binary attributes without retaining any of their significance. In addition to this, it provides us with a great concept of which characteristics are going to be the most significant for classification. The banking industry, the automobile industry, the healthcare industry, speech recognition, and the picture and text classification industries have all recently implemented this technique.

15.4 Result Analysis

Patients who were diagnosed with a variety of pancreatic tumors were included in the 500 CT scans of the abdomens that were included in the Medical Segmentation Decathlon (MSD) dataset [20]. When the model was put through its paces, it was tested with 100 photographs, while it was trained using 400 pictures as shown in Figure 15.3: Result Comparison of Machine Learning for Pancreatic cancer Detection.

15.5 Conclusion

Clinical diagnosis is difficult due to the lack of data source standardization. Improvements in the healthcare business have resulted in better diagnostic outcomes. According to clinical professionals, radiologists will be able to evaluate images and make diagnoses more accurately using computer-aided diagnosis technologies. There have always been several infectious diseases that kill humans. Pancreatic adenocarcinoma (PDAC) is a deadly disease

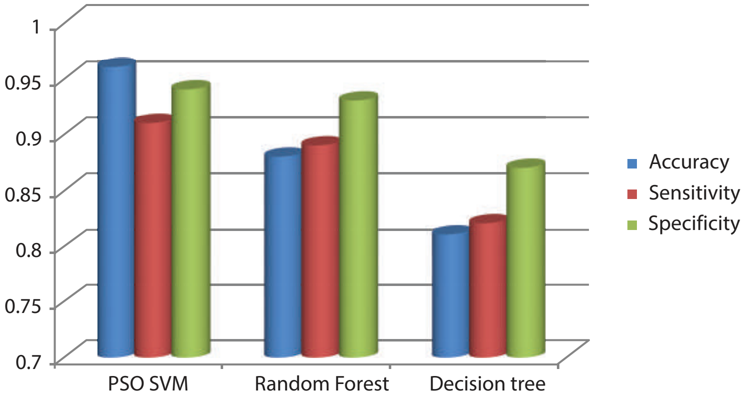


Figure 15.3 Result comparison of machine learning for pancreatic cancer detection.

that primarily affects adults over the age of 45. PDAC is the world's fourth most common cancer. Common diagnostic techniques include computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET) scans, and ultrasounds. An effective pancreatic cancer diagnostic system that uses multi-modal clinical data, such as structured EHRs, jaundiced eye images, and pancreatic CTs, can be constructed to forecast individuals' risk levels and detect pancreatic disease symptoms for rapid surgery planning. Pattern extraction from multi-modal healthcare data is difficult due to noise, the pancreas' shape and size, its placement in the abdomen, and a variety of other disease-related risk factors. The existing condition makes it extremely difficult to derive useful sickness patterns. The current segmentation approaches make it difficult to detect cancer in images of jaundice in the eyes or pancreatic CT scans. The goal of this research is to identify applicable disease patterns in clinical data to overcome the lack of a universal paradigm for diagnosing pancreatic ailment using multimodal clinical data and a variety of medical data sources.

References

1. Suman, G., Patra, A., Korfiatis, P. *et al.*, Quality gaps in public pancreas imaging datasets: Implications & challenges for AI applications. *Pancreatology*, 21, 5, 1001–1008, 2021.
2. Chen, P.T., Chang, D., Yen, H. *et al.*, Radiomic features at CT can distinguish pancreatic cancer from noncancerous pancreas. *Radiol. Imaging Cancer*, 3, 4, e210010, 2021.

3. Suman, G., Panda, A., Korfiatis, P., Goenka, A.H., Convolutional neural network for the detection of pancreatic cancer on CT scans. *Lancet Digit. Health*, 2, 9, e453, 2020.
4. Jasti, V., Zamani, A., Arumugam, K., Naved, M., Pallathadka, H., Sammy, F. *et al.*, Computational Technique Based on Machine Learning and Image Processing for Medical Image Analysis of Breast Cancer Diagnosis. *Secur. Commun. Netw.*, 2022, 1–7, 2022, doi: 10.1155/2022/1918379.
5. Ryan, D.P., Hong, T.S., Bardeesy, N., Pancreatic adenocarcinoma. *N. Engl. J. Med.*, 371, 1039–1049, 2014.
6. Kenner, B., Chari, S.T., Kelsen, D., Klimstra, D.S., Pandol, S.J., Rosenthal, M., Rustgi, A.K., Taylor, J.A., Yala, A., Abul-Husn, N. *et al.*, Artificial Intelligence and Early Detection of Pancreatic Cancer. *Pancreas*, 50, 251–279, 2021.
7. Yan, Z., Ma, C., Mo, J., Han, W., Lv, X., Chen, C., Chen, C., Nie, X., Rapid identification of benign and malignant pancreatic tumors using serum Raman spectroscopy combined with classification algorithms. *Optik*, 208, 1–10, 2020.
8. Savareh, B.A., Aghdaie, H.A., Behmanesh, A., Azadeh, B., Sadeghi, A., Zali, M., Shams, R., A machine learning approach identified a diagnostic model for pancreatic cancer through using circulating microRNA signatures. *Pancreatology*, 20, 6, 1195–1204, 2020.
9. Dayem Ullah, A.Z.M., Stasinos, K., Chelala, C., Kocher, H.M., Temporality of clinical factors associated with pancreatic cancer: a case-control study using linked electronic health records. *BMC Cancer*, 21, 1–13, 2021.
10. Hossain, S.M.M., Halsana, A.A., Khatun, L., Ray, S., Mukhopadhyay, A., Discovering key transcriptomic regulators in pancreatic ductal adenocarcinoma using Dirichlet process Gaussian mixture model. *Sci. Rep.*, 11, 1–15, 2021.
11. Jackins, V., Vimal, S., Kaliappan, M., Lee, M.Y., AI based smart prediction of clinical disease using random forest classifer and Naive Bayes. *J. Supercomput.*, 77, 5198–5219, 2021.
12. Li, Y., Zhang, J., Yuan, X., BagGMM: Calling copy number variation by bagging multiple Gaussian mixture models from tumor and matched normal next-generation sequencing data. *Digit. Signal Process.*, 88, 90–100, 2019.
13. Shastri, S., Kour, P., Kumar, S., Singh, K., Mansotra, V., GBoost: A novel Grading-AdaBoost ensemble approach for automatic identification of erythemato-squamous disease. *Int. J. Inf. Technol.*, 13, 959–971, 2021.
14. Pereira, S.P., Oldfield, L., Ney, A., Hart, P.A., Keane, M.G., Pandol, S.J., Li, D., Greenhalf, W., Jeon, C.Y., Koay, E.J., Almario, C.V., Halloran, C., Lennon, A.M., Costello, E., Early detection of pancreatic cancer. *Lancet Gastroenterol. Hepatol.*, 5, 7, 698–710, 2020.

15. Santosh Reddy, P. and Chandrasekar, M., PAD: A Pancreatic Cancer Detection based on Extracted Medical Data through Ensemble Methods in Machine Learning. *Int. J. Adv. Comput. Sci. Appl.*, 13, 2, 1–8, 2022.
16. Ahsan, M.M., Luna, S.A., Siddique, Z., Machine-Learning-Based Disease Diagnosis: A Comprehensive Review. *Healthcare*, 10, 3, 1–18, 2022.
17. Joloudari, J., Saadatfar, H., Dehzangi, A., Shamshirband, S., Computer-aided decision-making for predicting liver disease using PSO-based optimized SVM with feature selection. *Inf. Med. Unlocked*, *17*, 100255, 2019, doi: 10.1016/j.imu.2019.100255.
18. Lu, Y., Ye, T., Zheng, J., Decision Tree Algorithm in Machine Learning. *2022 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA)*, Dalian, China, pp. 1014–1017, 2022, doi: 10.1109/AEECA55500.2022.9918857.
19. Rajathi, V., Azeem A, S.A., S.S.P., Earlier Detection of Diabetes Using Random Forest Algorithm. *2022 1st International Conference on Computational Science and Technology (ICCST)*, Chennai, India, pp. 173–178, 2022, doi: 10.1109/ICCST55948.2022.10040471.
20. http://medicaldecathlon.com/
21. Rathi, M. and Rajavat, A., High Dimensional Data Processing in Privacy Preserving Data Mining, in: *2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT)*, pp. 212–217, IEEE, 2020, April.
22. Patsariya, M. and Rajavat, A., Network Path Capability Identification and Performance analysis of Mobile Ad hoc Network, in: *2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT)*, pp. 82–87, IEEE, 2020, April.
23. Dhar, S., Dhar, U., Rajavat, A., Factors and Attributes of Team Players: A Study of Engineering Students in India, in: *International Simulation and Gaming Association Conference*, pp. 53–60, Springer International Publishing, Cham, 2021, September.
24. Nahar, S., Pithawa, D., Bhardwaj, V., Rawat, R., Rawat, A., Pachlasiya, K., Quantum Technology for Military Applications, in: *Quantum Computing in Cybersecurity*, pp. 313–334, 2023.
25. Pithawa, D., Nahar, S., Bhardwaj, V., Rawat, R., Dronawat, R., Rawat, A., Quantum Computing Technological Design Along with Its Dark Side, in: *Quantum Computing in Cybersecurity*, pp. 295–312, 2023.
26. Chauhan, D., Singh, C., Rawat, R., Chouhan, M., Conversational AI Applications in Ed-Tech Industry: An Analysis of Its Impact and Potential in Education, in: *Conversational Artificial Intelligence*, pp. 411–433, 2024.
27. Chauhan, D., Singh, C., Rawat, R., Dhawan, M., Evaluating the Performance of Conversational AI Tools: A Comparative Analysis, in: *Conversational Artificial Intelligence*, pp. 385–409, 2024.

28. Shrivastava, M., Patil, R., Bhardwaj, V., Rawat, R., Telang, S., Rawat, A., Quantum Computing and Security Aspects of Attention-Based Visual Question Answering with Long Short-Term Memory, in: *Quantum Computing in Cybersecurity*, pp. 395–412, 2023.
29. Rawat, R., Chakrawarti, R.K., Raj, A.S.A., Mani, G., Chidambarathanu, K., Bhardwaj, R., Association rule learning for threat analysis using traffic analysis and packet filtering approach. *Int. J. Inf. Technol.*, *15*, 6, 3245–3255, 2023.

16

An Investigation of Various Text Mining Techniques

Rajashree Gadhave[1]*, Anita Chaudhari[2], B. Ramesh[3], Vijilius Helena Raj[4], H. Pal Thethi[5] and A. Ravitheja[6]

[1]*Computer Engineering, Pillai HOC College of Engineering and Technology, University of Mumbai, Mumbai, Maharashtra, India*
[2]*Mumbai University, Mumbai, India*
[3]*Institute of Aeronautical Engineering, Dundigal, Hyderabad, India*
[4]*Department of Applied Sciences, New Horizon College of Engineering, Bangalore, India*
[5]*Lovely Professional University, Phagwara, India*
[6]*Department of Civil Engineering, SVR Engineering College, Nandyala, Andhra Pradesh, India*

Abstract

Text summaries are generated from huge quantities of unstructured text, such as customer evaluations, internet log entries, and social media postings, and are an important feature of text mining systems. These synopses emphasize the key data representing the whole text or judgment. There are two major approaches to completing a summary task: extractive and abstractive. The extractive approach identifies the meat of the document, which contains all of the information, and extracts it to make the summary. In comparison to the other method, the abstractive methodology is more difficult since it creates summaries based on the document's keywords and semantics. The visually appealing summary job is challenging due to redundancy, a large text volume, unpredictability, and the semantics of natural language. Because of the task's complexity and breadth of applicability, researchers from the academia and business have been working hard on it. This study focuses on two major areas of text mining: feature-based text summarization and text similarity detection, both of which use machine learning-based methods.

**Corresponding author*: rgadhave@mes.ac.in

Rajesh Kumar Chakrawarti, Ranjana Sikarwar, Sanjaya Kumar Sarangi, Samson Arun Raj Albert Raj, Shweta Gupta, Krishnan Sakthidasan Sankaran and Romil Rawat (eds.) Natural Language Processing for Software Engineering, (233–244) © 2025 Scrivener Publishing LLC

Keywords**:** Text mining, clustering, classification, text summarization, text detection, machine learning

16.1 Introduction

There are a number of essential characteristics that should be considered when evaluating customer evaluations for the purpose of using them in decision-making for various commercial reasons. Most of the time, these evaluations appear to be disorganized and do not provide any hints about anything. For the purpose of achieving corporate goals, it is necessary to provide a concise summary of the most important aspects of these components. A great number of machine learning algorithms have previously rated and summarized the many elements available. The use of consumer preferences as a means of evaluating review components is given top priority in this study with the intention of assisting stakeholders in making more informed business decisions. When seen from the point of view of the customer, this reduces the difficulty of improving aspect ranking [1].

Methods of optimal and parallel large-scale analysis, in addition to machine learning, have been researched for their potential application in the process of summarization. This not only makes the summaries of a high quality, but also makes it possible to analyze extraordinarily large amounts of input from customers. Both machine learning and parallel approaches may be utilized to improve the performance of text summarization solutions. The current state of research in text mining is centered on issues concerning the representation of text, categorization of text, clustering of text, summarization of text, and modeling of hidden patterns [2].

The process of text mining involves the examination of vast quantities of unstructured material to obtain valuable insights. In the context of artificial intelligence, text mining and natural language processing may be viewed as techniques that assist in the extraction of meaningful information from excessive volumes of text. It is possible that this may be exploited for extra analysis, which would be advantageous for the operations of the establishment. There are a few different names for text mining, including data mining and analytics that apply to text. In our day and age, the internet has emerged as a significant source of data, and with it are enormous quantities of material that are not organized in any particular way. The great bulk of this unstructured data is derived from reviews that were posted by millions of different consumers. When these evaluations are subjected to quantitative analysis, companies may be able to arrive at more accurate conclusions. The identification of features and the summary of content that are useful are the two most fundamental reasons for the extensive usage of text mining and its applications. These are accomplishments that can be done using the procedures of feature extraction and text summarizing. By incorporating textual comments into these tasks, they are able to gain a deeper understanding of the operational challenges that are faced by potential customers. It is vital to increase the performance of automatic text summarization systems to deal with the large volumes of data that are created by the internet in the modern day [3].

Aspect ranking algorithms that have been developed lately take advantage of aggregation that is founded on domain ontology. Utilizing domain knowledge on its own will not be sufficient to make the system totally useable. It requires a higher level of precision in aspect ranking in accordance with the preferences of consumers and the knowledge of the area. The purpose of document or collection of input and text document summary is to generate a version of the original document or documents that are more condensed in length. At the moment, the bulk of approaches revolve around the concept of extracting certain words that contain descriptive information from a text. This summary is anything that is made up of a collection of these sentences. When it comes to extracting the semantic information of texts and applying it to build summaries, there are approaches that are more advanced than others [4].

The purpose of feature-based extractive summarizing algorithms is to search through customer reviews for terms that are relevant to the compilation of a summary that can be helpful. Nevertheless, in actuality, the vast majority of online assessments involve the use of tagging to uncover information on the domain in terms of emotion. Machine learning approaches, such as the parallel k-means clustering algorithm, have been utilized extensively by researchers to expedite the process of grouping and summarizing large datasets. This has been done to speed up the process. To solve the challenge of summarizing, Ferrari developed a method that made use of map-reduce, clustering, topic modeling, and semantic similarity. Both clustering and topic modeling are utilized by the system to summarize the texts.

Within the framework of the hybrid method, semantic data, knowledge models, and dictionaries are all included. Systems that are too intricate, such as this one, have repeatedly been unsuccessful. The amount of research that has been done on feature extraction and ranking algorithms that integrate domain knowledge with semantic information is insufficient. Better outcomes would be achieved with the use of this sort of system. The fact that it would be pointless to examine each and every one of these characteristics, the assignment that requires summarizing requires that they be ranked. It's possible that we may go over a number of things that don't provide anything of value to the experience of the customer [5].

A parallel process is utilized in the most recent development of systems for the construction of feature-based summaries. The concept of map-reduce is utilized in each and every one of these approaches to the processing and analysis of enormous volumes of data. Redundancy is yet another significant issue that arises with these systems. This thesis has a primary emphasis on aspect rating and summarization as its primary foci. The performance of existing aspect ranking systems, which have been created using graph-based and semantic techniques, could be improved by leveraging customer preferences as a basis for improvement. Although approaches that are based on semantic information have demonstrated some degree of promise, the effectiveness of these methods is contingent on the data that are contained within the content. The result of this is the development of a new ranking algorithm that takes into consideration the preferences of authors as well as domain knowledge that is generated from ontologies.

When it comes to the challenge of summarization, conventional approaches such as k-means clustering, optimization, and parallel processing lack the necessary effectiveness. At the core of these systems are issues that arise with regard to redundancy and scalability challenges. Because of this, three distinct methods of aspect summary have been proposed, which are as follows: In the first contribution, an implementation of a k-means method that combines optimization principles with distributed clustering is accomplished. A "bag of words" and real-valued vectors are examples of text representation models that are utilized in the second addition, which also makes use of the PSO approach. PSO's diversity enhancement and multi-objective functions are utilized in the development of this technique. An investigation on in-node optimization with combiner is included in the final plan. Through the utilization of map-reduce optimization, it is possible to scale even the largest datasets.

16.2 Related Work

In today's world, the availability of a wide range of text document sources has been made possible by computational automation. In the process of creating tools for text mining, one of the primary objectives is to organize the content of the text and identify patterns from among it. The process of text mining is comparable to that of data mining; however, text mining may handle data sets that are formless or semi-structured, whereas data mining methods are often utilized for the management of structured material [6].

Text mining is a subprocess of data mining that is widely used for the aim of extracting significant patterns and insights from huge volumes of unstructured textual data. It is also known as knowledge discovery. A wide range of methodologies from a variety of fields are applied in this text mining process. These methodologies include machine learning, visualization, statistics, knowledge management, natural language processing, information retrieval, case-based reasoning, database technology, and text analysis. Text mining has emerged as a fast- increasing field in the field of computer science [7]. This is due to the emergence of big data and artificial intelligence.

Text mining is a savior for computers that are attempting to validate unstructured computer data. This technique makes use of a wide variety of algorithms to transform free-form text into patterns that may be put into implementation. Text mining serves three primary functions: text summarization, text categorization, and text clustering. Each of these functions is important. An introduction to text mining, several approaches for text mining, and applications for text mining are all discussed in this study [8].

The practice of text mining is a branch of data mining. Data mining is the process of attempting to learn more by examining data. The technique of collecting relevant information from enormous volumes of text is referred to as text mining on the internet. In light of this, text mining is a process that includes extracting patterns from unstructured text [9]. This is analogous to the way that structured data is exploited in data mining.

Pattern discovery techniques provide the foundation for the majority of the methods that are used to find patterns. This strategy includes a number of different components, including data retrieval, topic extraction, clustering, and summarization [10]. For the purpose of information, extraction, an algorithm is aware of the key phrases and connections included inside a text. For the goal of accomplishing such, the pattern matching method is utilized. A comparison of the text that was entered by the user with the text sequences that were pre-defined is required to implement a pattern matching strategy. This approach is more effective for investigating huge text collections than other approaches. There is no simple method that can be utilized to transform the data that was retrieved into a format that is structured. In light of this, post-processing is an essential step [11]. The process of summarization is crucial to text mining. To put it in another way, summarizing is the process of delivering a concise overview of a lengthy document while maintaining the message or goal that is the primary focus of the text. It provides the user with assistance in judging the usefulness of the material. The process of compression is tied to summarization [12], despite the fact that it is not in a form that can be read by humans. Clustering does not make use of preexisting class labels; rather, it clusters items that are comparable together into groups according to the degree to which they are similar to something else. Few of the numerous categories that text clustering algorithms fall into include partitioning algorithms, agglomerative clustering algorithms, and fundamental parametric modeling-based techniques [13]. These are only few of the text clustering algorithms.

NLP stands for "natural language processing," which refers to the process of transforming human speech into a format that can be read by electronic devices. The basic objective of natural language processing [14] is to construct a computer system that is capable of analyzing, comprehending, and producing natural language processing. Among the various disciplines that make use of it, some examples include the translation of human languages, the creation of fictional worlds, and the development of robotic systems [15].

A KNN-based machine learning strategy was proposed by V. Bijalwan *et al.* [15] in 2014. This approach was proposed for the purpose of text and document mining. Data retrieval, also known as information retrieval (IR), is the process of looking for information in a variety of formats, including relational databases, documents, text, multimedia files, and the Internet. Internet search, automatic summarization, document clustering and classification, information filtering, image object extraction, online searching, and digital library searches are only some of the numerous applications of IR. Additional applications include the extraction of image objects. To categorize a particular query, the KNN (K-nearest neighbor) technique takes into account not just the document that is geographically closest to it, but also the categories of the k documents that are geographically closest to it. This is the fundamental premise that underpins the KNN approach. When this is taken into consideration, the Vector approach may be considered a subset of the KNN technique, where k is equal to 1. A vector-based, distance-weighted matching technique is utilized in this study, which is comparable to Yang's work in that it evaluates the degree of document similarity. Pictured below is a diagram that illustrates a schematic representation of the process of retrieving and indexing records. The KNN algorithm displayed the highest level of accuracy when compared to the Naive Bayes and Term-Graph algorithms. The temporal complexity of KNN is a significant drawback, despite the fact that it exceeds all of its competitors in terms of accuracy.

A hybrid feature selection technique for text classification that is based on an enhanced genetic algorithm was proposed by Abdullah Saeed Ghareb *et al.* [16] in the year 2016. This approach makes use of a hybrid search methodology that combines the advantages of filter feature selection techniques with an improved generalized algorithm (EGA) in a wrapper operation. This allows the method to concurrently improve classification performance while also addressing the large dimensionality of the feature space. To begin, we will discuss the EGA concept, which is founded on the operators for mutation and crossover. The crossover operation is carried out by splitting chromosomes (feature subsets) with session and document frequencies of chromosome entries (features). This is in contrast to the process of mutation, which is conducted by assessing the classifier execution of the original parents and the significance of features. Therefore, rather than depending on chance and random selection, the processes of mutation and crossover are carried out in accordance with the data that are pertinent. A concept for software that builds mind maps by utilizing a text mining method was proposed by Robert Kudeliu *et al.* [17] in the year 2011. The use of a mind map, which is based around a particular word or idea, allows for the visual representation of an assortment of things, including words, thoughts, projects, and other items. Mind maps may be applied for a variety of purposes, including but not limited to the following: research, organizing, problem solving, decision making, and writing. Additionally, they can be utilized for idea development, picturing, and classification. A graphic picture of a typical mind map is shown to us in reference number 119. The online page was downloaded in great detail to ensure that all of the essential information was there, which made this web data source appropriate.

The Rough Set Based Approach is one of the methods that Libiao Zhang *et al.* [18] proposed for the categorization of text. As a source of information on the internet, the gathering of textual materials has become increasingly important and significant. Text categorization is an important technology which plays a significant role in the management and organization of information. The fact that text categorization has become increasingly significant and relevant, researchers from all walks of life have developed an interest in the subject. In the beginning of this study, a variety of different feature selection approaches, algorithms for implementation, and applications of text categorization are presented. To enhance the performance of text categorization, there is a requirement for the development of additional novel strategies and methodologies. The reason for this is that the information that is retrieved by the data-mining algorithms that are currently being used for text classification has a sufficient amount of noise to provoke ambiguity in the process. The information extraction process and the practice of knowing both contribute to the creation of this ambiguity. Increasing the efficiency of the information extraction process and making effective use of the knowledge that has been gathered has been an essential step, but it has also been a considerable challenge. Utilizing a Rough Set decision-making process is advocated for the purpose of achieving more accurate document categorization in situations when standard text classification procedures are unsuccessful.

An approach that is based on fuzzy logic [18] was proposed for the purpose of text mining [19] and document clustering. The purpose of this article is to illustrate the application of fuzzy logic to text mining to cluster various documents. There is a possibility that the truth in fuzzy logic, which is a model of mathematical reasoning, might be biased. It can take on values between zero and one, which can be either completely true or completely false. Instead of relying on exact logic, it relies on approximations to make its decisions. It is the termination criteria that is utilized to decide when the iteration of the FCM algorithm comes to an end. The value of E can range anywhere from 0 to 1, and it is used to make this determination. To a lesser extent than E, the values of the fuzzy c-partition matrix have seen the most significant alteration.

The Huffman Encoding technique formed the foundation for the text clustering approach that was developed by Maria Muntean and colleagues [20] in the year 2014. An innovative strategy to enhance the accuracy of text data clustering is presented by the authors. Before expressing these attributes as numbers in the cluster assessment stage, our system encodes the textual contents of a dataset using the Huffman encoding technique. This is done before the cluster assessment step. We have achieved a significantly greater level of clustering accuracy when compared to the outcomes of more traditional methods. In the event that the dataset just contains string characteristics that require clustering, then this method will be effective. An in-depth explanation of the Huffman encoding algorithm.

16.3 Classification Techniques for Text Mining

16.3.1 Machine Learning Based Text Classification

Quantitative methods [21–24] for automating Natural Language Processing (NLP) with the help of machine learning algorithms make up Machine Learning based Text Classification (MLTC). The following material delineates the preferred supervised learning methods for text categorization. Classification methods range from the Rocchio algorithm, which relies on feedback, to instance-based learning algorithms [25, 26], which use both newly-stored and previously-trained instances; decision trees, support vector machines, and artificial neural networks—which include genetic algorithms and supervised learning algorithms—to name a few.

16.3.2 Ontology-Based Text Classification

By providing a framework for the formal definition of ideas, their descriptions, and the semantic links between them, ontology offers a potential answer to the issues at hand. There are several types of ontology, which stand for the meaning of data, including: domain area's ontology is its collection of concepts and the relationships between them. A few of the most fundamental building blocks of ontology include classes, attributes, relations, function words, and rules.

16.3.3 Hybrid Approaches

The text categorization was accomplished using many classification approaches. The hybrid approach produces effective results in text mining classification by combining several classification algorithms.

16.4 Conclusion

Text mining involves sifting through vast volumes of unstructured data to find relevant information. Artificial intelligence and natural language processing are frequently seen as methods for sifting through massive amounts of text in search of hidden insights. This study focuses on two significant subfields of text mining that employ machine learning techniques: feature-based text summarization and text similarity identification. Possible future research paths include creating summary systems that imitate the signals of all assessment metrics and investigating stakeholder feedback on summary quality using satisfactory indicators. Improving the system's efficiency and quality might also entail developing new algorithms for large-scale summarization. Text summarization systems assessed utilizing extrinsic methodologies might be useful for applications such as information retrieval and question answering.

References

1. Sukanya, M. and Biruntha, S., Techniques on Text Mining. *IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT)*, pp. 269–271, 2012.
2. Gupta, A. and Yadav, D., Semantic similarity measure using information content approach with depth for similarity calculation. *Int. J. Sci. Technol. Res.*, 3, 2, 56–59, 2014.

3. Paul, C., Rettinger, A., Mogadala, A., Efficient Graph-based Document Similarity. *International Sematic web Conference, the semantic web and latest advances*, pp. 17–25, 2016.
4. Derrac, *et al.*, A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm Evol. Comput.*, 1, 1, 3–18, 2017.
5. Gambhir, M. and Gupta, V., Recent automatic text summarization techniques: a survey. *Artif. Intell. Rev.*, 47, 1, 1–66, 2017.
6. Thilagavathi, K. and Priya, VS., Survey on text mining techniques. *Int. J. Res. Comput. Appl. Rob.*, 2014.
7. Kaushik, A. and Naithani, S., A Comprehensive Study of Text Mining Approach. *Int. J. Comput. Sci. Netw. Secur. (IJCSNS)*, 16, 2, 69, 2016.
8. Jadhav, A.M. and Gadekar, D.P., A Survey on Text Mining and Its Techniques. *Int. J. Sci. Res. (IJSR)*, 2012.
9. Rajawat, A. S., Rawat, R., Barhanpurkar, K., Shaw, R. N., & Ghosh, A., Vulnerability analysis at industrial internet of things platform on dark web network using computational intelligence. *Computationally intelligent systems and their applications*, 39–51, 2021.
10. Patel, F.N. and Soni, N.R., Text mining: A Brief survey. *Int. J. Adv. Comput. Res.*, 2, 4, 243–248, 2012.
11. Kanya, N. and Geetha, S., Information Extraction: A Text Mining Approach. *IETUK International Conference on Information and Comm. Technology in Electrical Sciences*, Dr. M.G.R. University, Chennai, Tamil Nadu, India, pp. 1111–1118, IEEE, 2007.
12. Gupta, V. and Lehal, G.S., A survey of text summarization extractive techniques. *J. Emerg. Technol. Web Intell.*, 2, 3, 258–268, 2010.
13. Agrawal, R. and Batra, M., A Detailed Study on Text Mining Techniques. *Int. J. Soft Comput. Eng. (IJSCE)*, 2, 6, January 2013.
14. Mishra, A. K., Tyagi, A. K., Dananjayan, S., Rajavat, A., Rawat, H., Rawat, A., Revolutionizing Government Operations: The Impact of Artificial Intelligence in Public Administration, in: *Conversational Artificial Intelligence*, pp. 607–634, 2024.
15. Bijalwan, V., Kumar, V., Kumari, P., Pascual, J., KNN based Machine Learning Approach for Text and Document Mining. *Int. J. Database Theory Appl.*, 7, 1, 2014.
16. Ghareb, A.S., Bakar, A.A., Hamdan, A.R., Hybrid feature selection based on enhanced genetic algorithm for text categorization. *Expert Syst. Appl.*, 49, 2016.
17. Kudeliu, R., Konecki, M., Malekoviu, M., Mind Map Generator Software Model with Text Mining Algorithm. *33 Int. Conf. on Information Technology Interfaces*, Cavtat, Croatia, June 27–30, 2011.
18. Zhang, L., Li, Y., Sun, C., Nadee, W., Rough Set Based Approach to Text Classification. *IEEE/WIC/ACM International Conferences on Web Intelligence (WI) and Intelligent Agent Technology (IAT)*, 2013.

19. Nahar, S., Pithawa, D., Bhardwaj, V., Rawat, R., Rawat, A., Pachlasiya, K., Quantum technology for military applications, in: *Quantum Computing in Cybersecurity*, pp. 313–334, 2023.
20. Sikarwar, R., Shakya, H.K., Kumar, A., Rawat, A., Advanced Security Solutions for Conversational AI, in: *Conversational Artificial Intelligence*, pp. 287–301, 2024.
21. Rathi, M. and Rajavat, A., High Dimensional Data Processing in Privacy Preserving Data Mining, in: *2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT)*, pp. 212–217, IEEE, 2020, April.
22. Patsariya, M. and Rajavat, A., Network Path Capability Identification and Performance analysis of Mobile Ad hoc Network, in: *2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT)*, pp. 82–87, IEEE, 2020, April.
23. Dhar, S., Dhar, U., Rajavat, A., Factors and Attributes of Team Players: A Study of Engineering Students in India, in: *International Simulation and Gaming Association Conference*, pp. 53–60, Springer International Publishing, Cham, 2021, September.
24. Pithawa, D., Nahar, S., Bhardwaj, V., Rawat, R., Dronawat, R., Rawat, A., Quantum Computing Technological Design Along with Its Dark Side, in: *Quantum Computing in Cybersecurity*, pp. 295–312, 2023.
25. Shrivastava, M., Patil, R., Bhardwaj, V., Rawat, R., Telang, S., Rawat, A., Quantum Computing and Security Aspects of Attention-Based Visual Question Answering with Long Short-Term Memory, in: *Quantum Computing in Cybersecurity*, pp. 395–412, 2023.
26. Rawat, R., Chakrawarti, R.K., Raj, A.S.A., Mani, G., Chidambarathanu, K., Bhardwaj, R., Association rule learning for threat analysis using traffic analysis and packet filtering approach. *Int. J. Inf. Technol.*, *15*, 6, 3245–3255, 2023.

17

Automated Query Processing Using Natural Language Processing

Divyanshu Sinha[1]*, G. Ravivarman[2], B. Rajalakshmi[3], V. Alekhya[4], Rajeev Sobti[5] and R. Udhayakumar[6]

[1]Department of Computer Science, MRIIRS Faridabad, India
[2]Department Of EEE, Karpagam Academy of Higher Education, Eachanari, Coimbatore, Tamil Nadu, India
[3]Department of Computer Science, New Horizon College of Engineering, Bangalore, India
[4]Institute of Aeronautical Engineering, Dundigal, Hyderabad, India
[5]Lovely Professional University, Phagwara, India
[6]Vel Tech Rangarajan Dr Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, Tamil Nadu, India

Abstract

Electronic papers are more accessible on a daily basis. Efficient data retrieval techniques are crucial for users to extract valuable information from vast quantities of electronic data, which has experienced significant growth in recent years. This article seeks to get insights from a Natural Language (NL) text query by examining several study areas in Natural Language Processing (NLP). Natural language processing (NLP) has experienced significant expansion in recent years, establishing itself as one of the most extensively studied subfields in artificial intelligence (AI). The researcher possessed a comprehensive understanding of natural language processing (NLP) techniques and their many applications, such as information retrieval (IR) and natural language interface to database (NIDB) systems. The Automated Intelligent Query Processing System (AIQPS) can handle text queries expressed in natural language. By employing several levels of natural language processing to the query, this goal can be accomplished. One of the challenges in natural language processing (NLP) is extracting knowledge from natural language text.

**Corresponding author*: divv4u@gmail.com
R. Udhayakumar: ORCID: 0000-0002-2447-664X

Rajesh Kumar Chakrawarti, Ranjana Sikarwar, Sanjaya Kumar Sarangi, Samson Arun Raj Albert Raj, Shweta Gupta, Krishnan Sakthidasan Sankaran and Romil Rawat (eds.) Natural Language Processing for Software Engineering, (245–258) © 2025 Scrivener Publishing LLC

***Keywords*: ** Natural language processing, intelligent query processing, artificial intelligence, SQL, information retrieval

17.1 Introduction

The pursuit of automating human languages is a long-standing goal within the field of Artificial Intelligence (AI). Intelligent Query Processing is centered around the utilization of several knowledge-base approaches from the AI field to automatically handle natural language text inquiries. For an intelligent system to understand, it is essential to possess wisdom, information, reasoning, cognition, language acquisition, and other related ideas. Intelligence is characterized by the acquisition, comprehension, and application of knowledge, as well as the ability to engage in abstract reasoning. To achieve the automation of intelligence, it is necessary to develop a computer system capable of doing tasks that typically require a significant level of intellectual capacity. It is possible to extract information from natural language documents by transforming them into a format that computers can process. Previously, academics in the domain of natural language processing (NLP) have attempted to manually assemble this information. Knowledge acquisition is a significant obstacle in the field of natural language processing (NLP) due to the need to encode a substantial amount of information that is both incomplete and erroneous. An estimated annual volume of NL text type information reaches around 160 terabytes. Academics are now exploring the intricacies of syntactic and semantic processing to unlock the potential advantages of natural language processing (NLP), which is known to be a complex task [1].

17.1.1 Natural Language Processing

The fundamental components of every language consist of its principles for manipulating symbols (grammar) and the set of symbols themselves (phonemes) that facilitate the communication of novel information through speech. Natural language processing (NLP) is concerned with constructing computational models capable of processing language in a manner similar to humans. This area requires a significant level of intellectual capacity. A computer may do natural language processing (NLP) when it has the skills of natural language understanding (NLU) and natural language generation (NLG). Natural Language Processing (NLP) refers to the theoretical framework that underlies computing techniques used to analyze text in natural language at various linguistic levels. A natural language (NL) document passes through one or many stages of language analysis, which is a thriving area of research in the field of artificial intelligence (AI). When individuals desire to express their ideas and emotions, Natural Language (NL) is the most widely used and direct system of symbols to employ. Natural languages and formal languages, used by computers, are distinct. Investigating natural language understanding (NLU), commonly referred to as computational linguistics, is crucial for enabling communication between computers and people. The cornerstone for the advancement of natural language understanding (NLU) lies in determining the research technique and direction, as well as describing and modeling natural language (NL). There are two primary paradigms in natural language processing: the data-driven approach and the knowledge-driven approach. Most of the earliest systems were founded on statistical principles. Various techniques have been proposed by researchers to enhance the performance of NLP systems [2].

The Annual Review of Information Science and Technology (ARIST) presents several theoretical breakthroughs, offering a comprehensive overview of the area. These encompass many techniques in natural language processing (NLP), such as statistical and corpus-based methodologies, research utilizing WordNet, computationally efficient finite-state approaches to NLP, and collaborative endeavors to create grammar and NLP tools. Statistical methods have several uses in natural language processing (NLP), such as word sense disambiguation, grammar construction and parsing, and others. They recommend use WordNet to augment the results of statistical analyses performed on texts produced from natural language. WordNet, developed at Princeton University, is an exceptional natural language processing (NLP) tool that categorizes verbs, adjectives, and adverbs into synsets, which are the fundamental units of lexical notions. To obtain further details on WordNet, one has to go to [Fellbaum, 1998] or conduct an online search. WordNet has been included into several research and applications in the field of natural language processing. Euro WordNet was established in 1996, marking a significant use of WordNet in natural language processing throughout Europe. Euro WordNet is structured in a manner that is comparable to the English WordNet. It consists of WordNets for other European languages, such as Dutch, Italian, Spanish, German, French, Czech, and Estonian [3].

Natural Language Processing (NLP) is an interdisciplinary field that combines linguistics and computer science to investigate the interaction between humans and computers. Computational linguistics and other disciplines of artificial intelligence should have significant overlap with language processing systems. The study of natural language processing (NLP) is necessary to explore the cognitive abilities associated with language due to its direct link to human cognition. The human mind responds to events because of its ability to engage in symbolic thinking, which entails managing intricate relationships and dealing with ambiguity. Natural Language Processing (NLP) provides a significant opportunity to apply statistical and logical approaches of Artificial Intelligence (AI). This study demonstrates the application of knowledge extraction techniques to get valuable insights from written texts in natural language. Comprehending the surroundings and the language itself is crucial for constructing computer models with the ability to comprehend human language. This, in turn, requires a grasp of how people acquire, retain, and process language.

17.2 The Challenges of NLP

The complexity of natural language processing (NLP) is influenced by several factors related to representation and interpretation. Representing NL accurately is difficult because of its significant level of ambiguity and vagueness. It is impossible to completely capture the extensive range and profound comprehension of human cognition in language processing. Compositional semantics considers the different meanings of words. Words alone do not form a whole phrase, as is often acknowledged. A language undergoes continuous evolution as new vocabulary is regularly incorporated and existing terms are repurposed in other situations. The terrorist attack on the World Trade Center in 2004 is commonly known as 9/11 among the majority of media outlets and television networks. When we process spoken or written language, there is a mental image that we can access. In the absence of access to explicitly coded comprehensive global or specific subject knowledge, a computer may alone comprehend the significance of a word in a message by analyzing its surrounding context. The co-occurring word context includes every word and its preceding or following elements. Deciphering written information gets increasingly challenging when idioms, metaphors, and ellipses are employed. Consider the following sentence: A rolling stone is incapable of collecting moss. Due to the fact that NL users rely on both explicit and implicit sources of information, they often fail to detect the ambiguity of the language. Effective language communication necessitates the involvement of both the speaker's and the reader's brains. We do not encode any information that the receiver should be aware of. To get a precise interpretation, the recipient must complete the missing information using the available resources. To understand the speech of others, those who acquire language must consider both past and present cultural knowledge and traditions. The incorporation of context and global information is the most significant obstacle to language processing. While those from outside of India may not instantly connect the name "Taj" with a hotel, landmark, or tea brand, it is a familiar association among Indians. Natural language (NL) has several sources of ambiguity at both the word-level and sentence-level. There are two potential sources of ambiguity for a word: "its" part of speech (POS) and its meaning. The word "can" has an ambiguous connotation in the present tense, although the word "bat" lacks a distinct meaning as well. We must develop several models and algorithms to address this issue. To determine if "can" functions as a noun or a verb, we may employ part-of-speech (POS) tagging. Similarly, to ascertain whether "bank" refers to a "financial institution" or a "river bank," we can utilize word-sense-disambiguation (WSD). The structural validation of a sentence in any natural language provides its unique linguistic characteristics, even if all natural languages have some common features such as syntax, alphabets, and lexicon. The main goal of computational linguistics is to establish a formalized framework for the grammar of natural language processing.

17.3 Related Work

It was recommended by Aghmazadeh and Dillig [1] that SQLizer, a system that is type-based and driven by database content, be used to generate SQLs from NLS database files. The structure in question developed its very own syntax, which may be thought of as a form of relational algebra. After that, it constructed a set of rules for the production of logical expressions that represent predicate clauses, and it utilized deductive rules to turn these expressions into a well-defined program in Structured Query Language. Methods such as rewriting and transformations based on inference rules were utilized extensively during the course of this research. It is not possible to create nested sub-queries using the language that Navid and Dillig [1] give since it is inadequate. A simple recommendation is made to regard nested SQLs as a sub-sketch of the main sketch, and it is suggested that the methods that are used to produce the initial query be repeated to construct the sub-queries. It is necessary to expand the language to build a complicated WHERE clause, which may include nested WHERE clauses. To summarize, SQLizer just offers a high-level description of nested query synthesize as part of its broader scheme of synthesis. It suggests that nested inquiries may be constructed by simply repeating the operation of the primary query, but it does not include instructions on how to actually make nested SQLs.

When it comes to handling the dynamic rearrangement and pipelining that is required when merging sub-queries, using rule-based rewriting techniques is not a clear way to address the situation. For the purpose of rating the intermediate logical forms or the initial sketch completions, SQLizer implements an internal scoring algorithm known as confidence scoring. This function uses a set of pre-defined features to determine confidence scores. High-quality training sets are either difficult to come by or complex to design, which makes it challenging to deploy NL2SQL systems that are based on machine learning. Rather than producing conventional (logic-based) outputs for a range of inputs, Deep Learning and Neural Network-based approaches are unable to handle complicated scenarios that may require dynamic interventions for rewriting, extension, or repair. This is because of the "black-box" nature of these techniques, which makes them incapable of handling such scenarios. Another category of NL2SQL systems is dependent on regular user input, which can be troublesome when it exceeds an acceptable limit, particularly when constructing complex SQLs. This is especially true when the limits are exceeded. It is necessary to make use of functional programming methodologies to provide the flexibility and composition that are required for the construction of nested inquiries. The fact that it offers a realistic viewpoint on automating query creation, it is ideally suited to the extensive collection of program transformation techniques that are included in Bird-Meetan's [2] formalism. The SQLizer specifier-repair approach is maintained in this work; however, it is expanded upon and a more in-depth examination of nesting is carried out. By applying more advanced methods for translation automation, particularly with respect to nested query integration, it goes above and beyond what is expected of it.

Linq, which stands for language integrated query, and other similar technologies are able to translate sentences written in imperative programming languages into SQL queries. The extent to which these systems are comprehensive and the degree of detail that they are able to manage (for example, nesting) relies on how effectively they function [3]. It is possible to convert embedded query expressions in programming languages into SQL instructions. These instructions may then be injected into the database to obtain the results that correspond to the expression that was provided in the programming language. Language was included and other programming languages have an impedance mismatch, which makes it impossible to perform recursion and other conventional operations on statements that might have unexpected implications in query expressions. This is because the impedance mismatch is caused by the conflict between the two languages. It is possible for translation to fail during runtime for a variety of reasons, such as run-time problems or a phenomenon known as partiality. Partiality happens when the translation fails to completely turn the expression into a SQL query and instead creates just a partial sketch. Without even taking into consideration nested query structures, this is already a scenario that is problematic. Although it is well known that the FERRY [4] system offers a comprehensive translation, it is important to note that it does not offer any support for abstraction.

Examples of non-linear information databases (NLIDBs) that make use of ontologies in combination with an easy-to-use authoring module are provided in [5, 6]. When the DRS keywords are contained within a query context, the query context is a subgraph of the semantic graph. This particular subgraph is built on the foundation of associations between tables. The information that is obtained from NLQ/DRS is then turned into data items before being stored in the database. These data items include the names of tables and views, relations, attributes (column names), and values. To produce the final SQL query, the query context, select list, and relational constraint columns are employed. The select clause may contain any columns that are not part of a constraint and are indicated in the NL query. These columns are considered potential inclusions.

This is accomplished through the utilization of ontology-assisted programming paradigms, which allow natural language programming to make use of entities derived from natural language phrases. A natural-language user interface is yet another tool that is developed with the intention of simplifying the translation process for users. The process of automatically producing programs that are compliant with a predetermined collection of input-output examples has also been the subject of research. The implementation of tree transformations on Hierarchical Data Trees was also accomplished through the utilization of systems that generate path transformation functions in the intermediate language. The very first graphical query language was called Query-By-Example (QBE), and it was used to list and show database tables. It also gave the user the ability to input instructions, example items, and criteria that were centered on these tables to obtain data. One of the fundamental premises of QBE is that the input text is converted into DML instructions (like SQL) by utilizing parser. It is possible for skeleton tables in QBE to be connected to actual tables in the database. This means that whenever the user performs an action on the database, an example of a solution to that operation is filled in. SEQ2TREE [7] is a Neural Network model that transfers NL statements into their semantic representations. This translation is accomplished with the assistance of an additional attention layer. NL utterances are converted to vectors by SEQ2Tree, which then builds sequences or trees that reflect the logical forms of the vectors. This process is carried out by RNNs using LSTM units. It is necessary to encode the NL tokens at the beginning as a vector for the neural network to be able to process an NLQ initially. With regard to the majority of cases, attention models are also necessary to improve translation accuracy. In comparison to the SEQ2SEQ model, the SEQ2TREE model is superior in terms of performance since it clearly models the compositional structure of logical forms. This is accomplished by including a Hierarchical Tree (H-Tree) decoder. Within sequence-to-sequence decoders, the production of tokens is restricted because it is dependent on all tokens that were formed before it. In lieu of semantic parsing of aggregations and fundamental WHERE clause requirements that justify the development of nested predicate clauses, all Sequence-to-Sequence models disregard the possibility of producing nested Sub-Queries. This is the case regardless of whether or not Reinforcement Learning is utilized. In the event that an NL query and the schema of a database are provided, the Neural Network model SQLNet [8] is able to generate SQL queries. This is accomplished by utilizing the Seq2SQL architecture. The Initial Query Sketch is used by SQLNet to offer the dependency relationship of distinct slots in the sketch. This is done to guarantee that the predictions of each slot are in agreement with the forecasts of its dependent slots. SQLNet presents two new constructs—sequence-to-set and column attention—to put this concept into action. These constructions are derived from the attention map that is constructed by utilizing the properties of the NL questions and the names of the columns in the database table. Sequence-to-set eliminates the need for reinforcement learning and makes it possible to repair by rewriting predicate clauses. This is accomplished by eliminating the constraint that "order does not actually matter" from the generation of predicate clauses. Deep learning systems such as SEQ2SQL [8] and SQLNet [9] have only been examined on datasets that included simple NLQs. This is the only dataset that has been used for evaluation. As a result of the fact that SEQ2SQL and SQLNet are restricted to the creation of SQL queries over a single table, nested SQL queries are not a good fit for these two databases. For the purpose of integrating many tables through the utilization of JOIN clauses, neither SQLNet nor Seq2SQL performs exceptionally well, despite the fact that SQLNet offers greater accuracy.

In April of 2018, P. Utama and others released a paper with the title "DBPal: An End-to-end Neural Natural Language Interface for Databases" [10]. An auto-completion model that was learned was applied by the authors to compose complicated queries. Additionally, Deep Neural Networks, more especially a sequence-to-sequence Recurrent Neural Network model, were utilized to convert NLS to SQL. Utilizing the database structure in conjunction with a collection of basic templates and "Slot filling" dictionaries, DBPal generates a tedious training set consisting of one to two million pairings to acquire every possible NL-SQL pair. This is done to acquire every possible pair. Within the WHERE clause, DBPal only generates Sub-Queries that are not connected with one another. The Query Sketches that are developed for the purpose of conventional NL to SQL translation are enlarged for the building of nested queries, just as they are expanded for the development of non-nested queries. With regard to 'patients,' DBPal functions more effectively than Geo DB.

When Iyer *et al.* [11] were putting together their implementation of an interactive user-based feedback learning scheme, they were required to manually construct a training set for each and every database and schema. This was done to build NL to SQL pairings that were a match. The Neural Semantic Parser (NSP) that they have developed is a model that directly translates natural language queries (NL) into SQL queries by utilizing the capabilities of the target database. This encoder-decoder paradigm is powered by a bidirectional LSTM network, which is responsible for international attention. Following the receipt of binary input from users on its predictions to lower and the set of queries that require labeling, NSP makes use of paraphrasing to reduce the amount of annotation efforts, which ultimately leads to more efficient learning. When it comes to performance, this paradigm can only compete with systems that directly construct SQL from NLS or those that translate speech into logical representations and then synthesize SQLs. These are the only systems that can effectively compete with this paradigm because it depends on SQL templates to generate the final SQL query, NSP is not ideal for circumstances that need the development of nested or co-related sub-queries. This is because NSP is less dynamic than other SQL queries because it relies on SQL templates.

17.4 Natural Language Interfaces Systems

An interface allows any search engine (IR system) to receive a user's request. The search engine will employ it and convert it into requests. Typically, the translation will include a compilation of index phrases or keywords that succinctly explain the queries. Data retrieval systems search structured persistent storage for things that satisfy specific conditions, such as Regular Expressions (RE) and Relational Algebraic (RA) Expressions NL. The textual data that an information retrieval system operates on is inherently lacking in organization and may include uncertainty in its semantic interpretation. By accepting an NL query, this study can reveal the complexities of user interaction with the system. Figure 17.1 depicts the Natural Language Query Processing (NLQP) in both the Natural Language Processing (NLP) application areas. The user submits a natural language query to a question processing system to obtain keywords. The system will evaluate the query using natural language processing techniques. In the next step, the keywords are inputted into the information retrieval (IR) system as a query to obtain articles that are rated based on their relevance. When dealing with structured databases, natural language queries can be turned into relational algebra (RA) or relational expressions (RE), which can then be easily transformed into SQL queries. The outcomes will be delivered in a format that adheres to the principles of a relational database as shown in Figure 17.1: Natural language query processing.

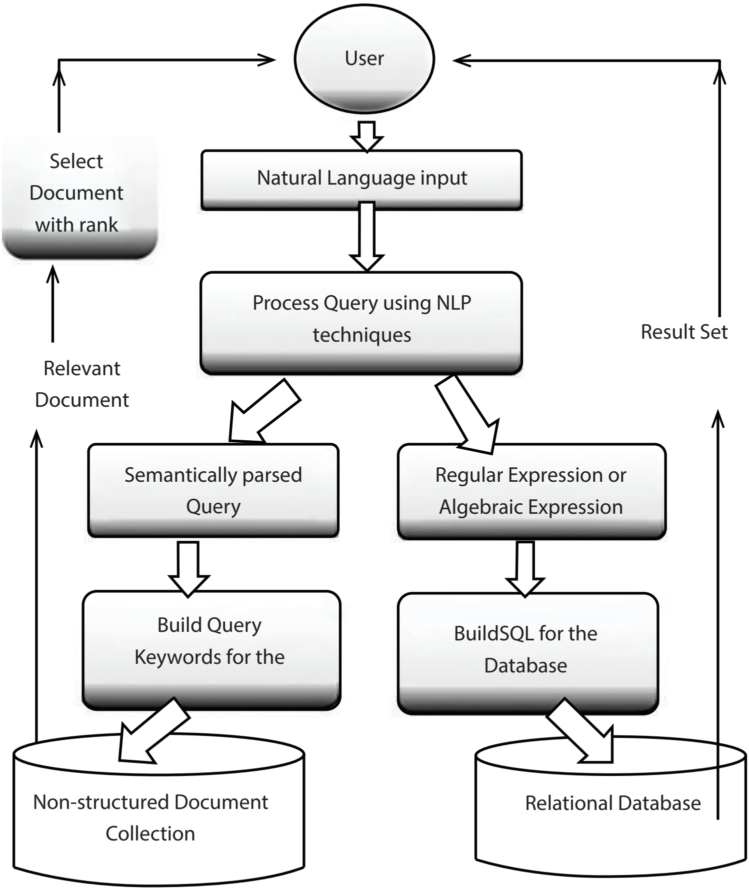

User
Natural Language input
Process Query using NLP techniques
Select Document with rank
Relevant Document
Result Set
Semantically parsed Query
Regular Expression or Algebraic Expression
Build Query Keywords for the
BuildSQL for the Database
Non-structured Document Collection
Relational Database

Figure 17.1 Natural language query processing.

Artificial intelligence researchers face a daunting task of developing systems that can accurately replicate human intelligence. The ability to engage in metaphorical thinking is fundamental to how the human mind processes different situations. Intelligence refers to the capacity to engage in cognitive processes such as thinking, reasoning, and comprehending intricate concepts. Individuals assimilate new information into their cognitive faculties as knowledge. The human mind [11–14] represents information as symbols, which may be altered through the process of thinking. AI [15–17] researchers face several challenges, one of which is the creation of a computer that can understand human knowledge. Knowledge representation is a crucial aspect of AI. Language may be used to convey human mind, especially complex and consciously inferred thoughts. The integration of artificially intelligent systems poses significant difficulties in the creation of NL-aware applications [15, 18]. Knowledge serves as the fundamental basis for many forms of logical thinking and behavior, encompassing problem-solving, task delineation, inference, and decision-making. Research on knowledge representation is propelling the advancement of the information age, transitioning it from its initial stage of data processing [16, 19, 20] to a more sophisticated stage of knowledge processing. However, there are currently alternative methods for encoding knowledge, including production rules, logic, script, semantic networks, and frames. An urgent concern in the field of artificial intelligence now revolves around the necessity to uncover more effective methods for representing knowledge. Fortunately, knowledge graphs are an innovative method that expands upon the methodology of knowledge representation. This approach utilizes principles from philosophy and psychology to construct a semantic representation model that elucidates the way in which humans perceive and process information. This thesis presents the author's research on natural language processing (NLP), intelligent query processing (IQP), and machine learning (ML). Both sectors have conducted thorough and comprehensive research; hence, merging them will enable us to use their individual capabilities.

17.5 Conclusion

It is the goal of this research to find a solution to the problem of intelligently processing a natural language text inquiry by utilizing natural language processing (NLP). Examining the many different query languages that a normal user may use to ask a question is done to determine the level of difficulty associated with query processing. This article presents a structured framework for the problem of inquiry processing, as well as a number of intelligent approaches from the literature on artificial intelligence that have the potential to manage inquiries. These methods include the Conceptual Graph, the Ontology, and the Frame techniques. The utilization of natural language is among the most ingenious methods of expressing one's expertise.

References

1. Mishra, A. K., Tyagi, A. K., Dananjayan, S., Rajavat, A., Rawat, H., Rawat, A., Revolutionizing Government Operations: The Impact of Artificial Intelligence in Public Administration, in: *Conversational Artificial Intelligence*, pp. 607–634, 2024.
2. Bird, R.S., An Introduction to the Theory of Lists, in: *Logic of Programming and Calculi of Design. NATO ASI Series*, vol. 36, M. Broy (Ed.), pp. 5–42, Springer Verlag, USA, 1987.
3. Bhalotia, G., Hulgeri, A., Nakhe, C., Chakrabarti, S., Sudarshan, S., Keyword searching and browsing in databases using BANKS, in: *ICDE*, 2002.
4. Grust, T., Mayr, M., Rittinger, J., Schreiber, T., FERRY: Database-supported program execution. *SIGMOD-PODS'09 – Proc. Int. Conf. Manag. Data 28th Symp. Princ. Database Syst*, pp. 1063–1065, 2009.
5. Lopez, V., Pasin, M., Motta, E., Aqualog: An ontology-portable question answering system for the semantic Web, in: *The Semantic Web: Research and Applications*, 1st ed., vol. 3532 of Lecture QNotes in Computer Science, A. Gómez-Pérez and J. Euzenat (Eds.), pp. 546–562, Springer-Verlag Berlin Heidelberg, Crete, Greece, 2005.
6. Bernstein, A., Kaufmann, E., Göhring, A., Kiefer, C., Querying ontologies: A controlledenglish interface for end-users, in: *The Semantic Web - International Semantic Web Conference*, vol. 3729 of Information Systems and Applications, Y. Gil, E. Motta, V.R. Benjamins, M. Musen (Eds.), pp. 112–126, Springer, Galway, Ireland, 2005.
7. Nahar, S., Pithawa, D., Bhardwaj, V., Rawat, R., Rawat, A., Pachlasiya, K., Quantum technology for military applications, in: *Quantum Computing in Cybersecurity*, pp. 313–334, 2023.
8. Sikarwar, R., Shakya, H. K., Kumar, A., Rawat, A., Advanced Security Solutions for Conversational AI, in: *Conversational Artificial Intelligence*, pp. 287–301, 2024.
9. Pithawa, D., Nahar, S., Bhardwaj, V., Rawat, R., Dronawat, R., Rawat, A., Quantum Computing Technological Design Along with Its Dark Side, in: *Quantum Computing in Cybersecurity*, pp. 295–312, 2023.
10. Namdev, A., Patni, D., Dhaliwal, B. K., Parihar, S., Telang, S., Rawat, A., Potential Threats and Ethical Risks of Quantum Computing, in: *Quantum Computing in Cybersecurity*, pp. 335–352, 2023.

11. Shrivastava, M., Patil, R., Bhardwaj, V., Rawat, R., Telang, S., Rawat, A., Quantum Computing and Security Aspects of Attention-Based Visual Question Answering with Long Short-Term Memory, in: *Quantum Computing in Cybersecurity*, pp. 395–412, 2023.
12. Rathi, M. and Rajavat, A., High Dimensional Data Processing in Privacy Preserving Data Mining, in: *2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT)*, pp. 212–217, IEEE, 2020, April.
13. Patsariya, M. and Rajavat, A., Network Path Capability Identification and Performance analysis of Mobile Ad hoc Network, in: *2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT)*, pp. 82–87, IEEE, 2020, April.
14. Dhar, S., Dhar, U., Rajavat, A., Factors and Attributes of Team Players: A Study of Engineering Students in India, in: *International Simulation and Gaming Association Conference*, pp. 53–60, Springer International Publishing, Cham, 2021, September.
15. Nahar, S., Pithawa, D., Bhardwaj, V., Rawat, R., Rawat, A., Pachlasiya, K., Quantum Technology for Military Applications, in: *Quantum Computing in Cybersecurity*, pp. 313–334, 2023.
16. Pithawa, D., Nahar, S., Bhardwaj, V., Rawat, R., Dronawat, R., Rawat, A., Quantum Computing Technological Design Along with Its Dark Side, in: *Quantum Computing in Cybersecurity*, pp. 295–312, 2023.
17. Chauhan, D., Singh, C., Rawat, R., Chouhan, M., Conversational AI Applications in Ed-Tech Industry: An Analysis of Its Impact and Potential in Education, in: *Conversational Artificial Intelligence*, pp. 411–433, 2024.
18. Chauhan, D., Singh, C., Rawat, R., Dhawan, M., Evaluating the Performance of Conversational AI Tools: A Comparative Analysis, in: *Conversational Artificial Intelligence*, pp. 385–409, 2024.
19. Shrivastava, M., Patil, R., Bhardwaj, V., Rawat, R., Telang, S., Rawat, A., Quantum Computing and Security Aspects of Attention-Based Visual Question Answering with Long Short-Term Memory, in: *Quantum Computing in Cybersecurity*, pp. 395–412, 2023.
20. Rawat, R., Chakrawarti, R.K., Raj, A.S.A., Mani, G., Chidambarathanu, K., Bhardwaj, R., Association rule learning for threat analysis using traffic analysis and packet filtering approach. *Int. J. Inf. Technol.*, *15*, 6, 3245–3255, 2023.

18

Data Mining Techniques for Web Usage Mining

Navdeep Kumar Chopra[1]*, Chinnem Rama Mohan[2], Snehal Dipak Chaudhary[3], Manisha Kasar[4], Trupti Suryawanshi[5] and Shikha Dubey[6]

[1]Department of CSE, Seth Jai Parkash Mukand Lal Institute of Engineering and Technology, JMIT, Radaur, USA
[2]Department of Computer Science and Engineering, Narayana Engineering College, Nellore, Andhra Pradesh, India
[3]Information Technology, Bharati Vidyapeeth (Deemed to be University) College of Engineering Pune, Maharashtra, India
[4]Computer Engineering, Bharti Vidyapeeth (Deemed to be University) College of Engineering, Pune, Bharti Vidyapeeth Deemed to be University, Pune, Maharashtra, India
[5]Computer Science and Engineering, Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune, Bharati Deemed to be University, Pune, Maharashtra, India
[6]MCA, Dr. D. Y. Patil Institute of Management and Research, Pimpri, USA

Abstract

Web use mining is the application of data mining and computational methods to analyze and understand web server logs, which provide insights on how individuals utilize the internet. The crucial data requirements of the computational methods are fulfilled by obtaining and converting web server logs. Computational algorithms aid in constructing learning models that enhance the online environment by efficiently detecting interesting patterns. The primary areas of emphasis in web research are web cache optimization and online customization. This research aims to optimize web caching and enhance web personalization by analyzing significant online trends, including user groups and categories that engage with interesting content at a later time. Improved web cache design may be achieved by

***Corresponding author*: navdeepkumar17@gmail.com; ORCID: 0000-0003-4751-0326
Chinnem Rama Mohan: ORCID: https://orcid.org/0000-0001-9209-3029

Rajesh Kumar Chakrawarti, Ranjana Sikarwar, Sanjaya Kumar Sarangi, Samson Arun Raj Albert Raj, Shweta Gupta, Krishnan Sakthidasan Sankaran and Romil Rawat (eds.) Natural Language Processing for Software Engineering, (259–270) © 2025 Scrivener Publishing LLC

using predictive algorithms that anticipate the categories of interesting pages that will be browsed next. Data about user interests obtained via the creation of groups may be used to customize the online experience for individual visitors. This paper extensively explores the subject of online content mining and web usage mining. In this chapter, we introduce decision tree techniques for web usage mining.

***Keywords*:** Web content mining, web usage mining, data mining, data preprocessing, decision tree, pattern analysis

18.1 Introduction

Web miners may efficiently and automatically extract useful information and discover interesting patterns by using information mining methods on data stored on the World Wide Web. Although web mining has its origins in the discipline of information mining, it is important to note that the two are separate and different. The unstructured nature of online data is an intrinsic characteristic that adds to the unpredictable nature of web mining. Web mining is the systematic extraction of extensive information on a user's interactions with a website via the analysis of their server logs, which are also referred to as web server logs. Almost all online projects may considerably benefit from the valuable information obtained via an analysis of server logs [1].

Web mining may be classified into three primary categories. One area is web content mining, which focuses on extracting information from internet sites, including both text and visuals. Web structure mining is the process of collecting data or information from the structures of web pages and analyzing how these structures impact navigation. This helps in comprehending the user's requirements and integrating them into online applications. Online use mining is the process of deriving valuable usage patterns from online data. It encompasses several crucial applications such as online customization, site design, e-commerce, advertising, marketing, fraud detection, transaction analysis, and educational data mining [2].

18.1.1 Web Usage Mining

The primary objective of web use mining is to get more insights on the behavior of online users and their interactions with websites. Web use mining seeks to efficiently and effectively identify patterns of user access from web log data. Web usage mining enables the retrieval of data pertaining to client access, including server logs, registration information, and other relevant details. Web mining steps are discussed as follows:

1) Data Collection

Various types of online log information document storage can be accessed to successfully retrieve log information.

A) Web server logs (server-side log data): The web server stores its logs internally. Web server logs may include various information, such an IP address, requested URL, time stamp, bytes, and the convention used. Typically, this data [15, 16] is presented in a certain manner. The Common Log Format (CLF) is widely used as one of the most common log record formats. A web server generates a standardized log design document to monitor and record the requests made on a website.

B) Proxy server logs (proxy side log data): The data that is saved on the intermediary server is referred to as proxy server logs. During periods of temporary unavailability of the main server, the intermediate servers take up the responsibility of handling client requests. Consequently, the intermediary servers create logs [17, 18]. Intermediate server logs provide further information related to the intermediary server, in contrast to web server log records.

C) Browser logs (client-side log data): The client's own computer, from which they access websites, has its browser records collected. Developers use Java or JavaScript to deploy remote agents that are then integrated into web sites to get client data, including a user's browsing history. Disregarding problems [19, 20] such as caching and IP misinterpretation, data obtained from clients are more reliable than data obtained from servers. Users must provide help to access these data. However, users often restrict the capability of Java and JavaScript apps due to security [18, 19, 21–23] concerns.

2) Data Pre-Processing

Data pre-processing is a crucial step in web use mining. After collecting vast amounts of data from various sources, the subsequent stage involves data preparation. Prior to proceeding to pattern identification, it is essential that the data be both consistent and integrated. Data preparation include information cleaning, client identification, session identification, and pathway completion [3].

3) Pattern discovery

A) Association: The technique relies on the establishment of a consistent pattern and the formulation of rules as its fundamental basis. Upon doing initial processing, the data extracted from the web log file uncovers several captivating insights, such as the ratio of URL visits to users. This information enables the identification of frequently visited websites; hence aiding in the comprehension of user preferences and wants. The main objective of the association rule is to identify relationships among web pages that have been seen by the user. An association rule may be used to link a user's most frequently visited landing page to their session. Association rule mining utilizes several techniques, including Apriorism, Eclat, Frequent Pattern tree, and others [4].

B) Clustering: Clustering is a technique used to group together persons or data elements that have similar features or characteristics. It has the potential to be a valuable advantage in the creation and execution of next marketing plans. Users may be clustered together based on their similar navigational behaviors to establish groups. Utilizing market dispersion in e-commerce operations or delivering personalized online content to targeted customers may be significantly enhanced by deducing user data. Internet search engines and Web service providers profit from page clustering since it enables them to identify groupings of pages that contain relevant material [5].

C) Sequential pattern: There are many techniques, such as Apriorism, SPADE, GSP, Prefix Span, and Spam, that may be used for sequential pattern analysis. The objective of this investigation is to ascertain the temporal gap between the visits of a suspected user to links X and Y. This research has the potential to enhance criminal detection by uncovering the suspected user psychology [6].

D) Classification: This technique allocates each data element to one of many pre-established categories, which then function as the foundational components for user profiles. Initially, a person will be required to extract and choose features that distinctly describe the attributes of a certain class or category. Grouping may be achieved by the use of directed inductive learning techniques such as Support Vector Machines, k-nearest neighbor classifiers, naïve Bayesian classifiers, and decision tree classifiers [7].

4) Pattern Analysis: During the pattern analysis phase, a valuable model or standard pattern is discovered for a certain internet use mining application. Pattern analysis employs several techniques such as visualization, online analytical processing (OLAP), data and knowledge querying, and usability analysis.

18.2 Web Mining

Web mining is a specialized kind of data mining that focuses on extracting valuable information from websites and other online data sets by analyzing user habits and structural components. A diagram (Figure 18.1) illustrating the categorization of web mining is available.

There are three methods which are relevant for web mining. Web Content Mining• Web Structure Mining• Web Usage Mining

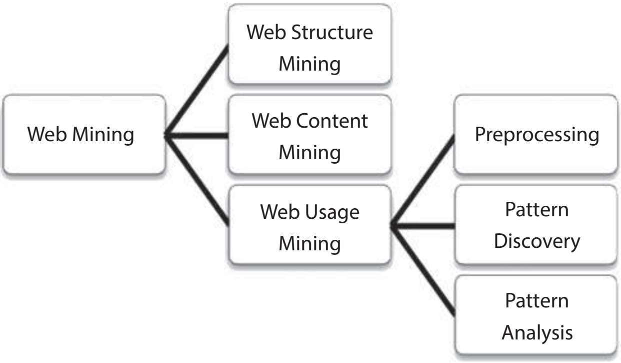


Figure 18.1 Steps involved in web mining.

18.2.1 Web Content Mining

In essence, web mining refers to the extraction of data from the internet. There are necessary procedures for retrieving the data saved on the internet. The activity is referred to as online content mining. The internet provides a vast amount of information, with many websites accessible to users. The web content includes activities such as information retrieval and accessing search engine sites. The content mining result pages provide the most recent and precise result [8]. The various contents of Web Content Mining are -

Web Page: A web page often contains a variety of information, including core content, advertisements, navigation panels, copyright notices, and more. For every particular job, only a portion of the data is useful and can be acted upon, while the rest is just irrelevant information.

Search Page: A frequent use of a search page is to do several searches inside a certain webpage of a website. A content database categorizes and groups web pages in a manner that facilitates easy navigation for both users and search engines.

Result page: Content mining result pages often include the search results, the visited websites, and an explanation of the most recent accurate result.

18.2.2 Web Structure Mining

It is feasible to define web structure mining using a graph-based approach. Web pages may be seen as nodes, while hyperlinks can be likened to edges in a network diagram [9]. The essential link between users and the web has been unveiled. The objective of web structure mining is to provide systematic synopses of material discovered on the internet. Below is a hyperlink that connects one website to another. The various contents of Web structure mining are -

Links Structure Mining: Link analysis is a well-established area of research. However, with the increasing popularity of Web mining, researchers delve further into structural analysis, giving rise to a new discipline known as link mining. Classification, clustering, connection type, link strength, and link cardinality are all components of it.

Internal Structure Mining: It enhances search results via the process of filtering. Additionally, it endeavors to reveal the underlying architecture of the internet's link structures and may provide insights into page ranking or authority. By using this methodology, the interconnections and resemblances across different websites can be analyzed.

URL Mining: A hyperlink is a fundamental element that allows for navigation across online pages, including the ability to redirect to a completely other web page.

18.2.3 Web Usage Mining

Web use mining [10] enhances overall speed for future accesses by facilitating the identification of essential links and prioritizing the search process via the recognition of users' frequent access habits. Web Usage Mining is the use of data mining techniques to identify trends in online data, with the goal of gaining a deeper understanding of web-based services and meeting their specific needs. Web Usage Mining is classified into three aspects,

- Preprocessing
- Pattern Discovery
- Pattern Analysis

The main focus of the Web Usage Mining technique is to analyze the data stored in log records. The registration folder of a website comprises logical data pertaining to consumer requirements. One potential objective of online use mining is to provide useful recommendations to website visitors or to enhance website organization. The main objective of internet use mining is to extract significant patterns from extensive datasets, including visually captivating or otherwise fascinating patterns.

18.2.3.1 Preprocessing

A custom tread that incorporates data breakdown from these use logs as well as integration with other use logs is a data-focused approach [11]. Finding out what kind of file includes a prohibition, such as a hyperlink with text, helps someone to finish this procedure. Data cleaning is the process of identifying and removing errors and inconsistencies to improve data quality.

18.2.3.2 Pattern Discovery

The analyst utilizes the obtained data to assist the organization in formulating marketing strategies, and one method to accomplish this is by identifying patterns. It is also important to understand the key vocabulary used in web mining. After session identification, the organization has the freedom to select any data mining technique, such as classification analysis, association rule discovery, sequential pattern discovery, clustering analysis, or statistical analysis, based on the analyst's needs. The utilization of classification analysis is performed to categorize customer profiles. The process of association rule discovery in web mining enables the identification of the most frequently accessed pages during user sessions. By analyzing data from a collection of records related to frequently visited web pages, it is feasible to enhance website design and predict customer behavior in connection to the most often seen sites. Sequential pattern identification is a significant advancement in the field of internet use mining. It may be used to predict client behavior by collecting data from web logs that record their surfing patterns. Offering personalized attention to clients is a strategy that organizations may employ to enhance connections with new customers and retain existing ones.

18.2.3.3 Pattern Analysis

Pattern analysis is the final stage in the process of extracting insights from online usage data. After the discovery of patterns, they undergo thorough evaluation to exclude any extraneous information from the gathered data. The organization highly prioritizes this phase since it enables them to easily comprehend client behavior. Pattern analysis employs several methodologies such as visualization, online analytical processing (OLAP), data and knowledge querying, and usability testing.

18.3 Web Usage Data Mining Techniques

The various techniques are used to acquire the hidden information in Web usage mining.

A. Decision trees
A decision tree is composed of rectangular forms that represent the core nodes and oval shapes that represent the leaf nodes. This structure gives the decision tree a resemblance to a flowchart. Each internal node has at least two child nodes. Split nodes, which assess the worth of an attribute expression, are found in every internal node. Every test result is linked to an arc that stretches from a parent node to its children. Each leaf node is assigned a class label. Decision trees are often used as a method for collecting data for making decisions. Each decision tree has a primary node where the user may execute actions. From this node, users iteratively partition each

node according to the decision tree learning process. The final outcome is a decision tree, where each node represents a possible choice and its corresponding result. The ID3, CART, CHAID, and C4.5 decision tree learning algorithms are the four most often used ones [12].

THE ID3 DECISION TREE

The ID3 method, devised by J.R. Quinlan, selects successive attributes based on their associated information gain. The test characteristic for the current node is chosen based on the one that offers the most information gain or decrease in entropy. Ross Quinlan developed ID3, a direct method for acquiring decision trees. The ID3 approach is based on a top-down, greedy search that evaluates each characteristic at every node of the decision tree using the given sets. A metric called "information gain" is presented to determine the most beneficial attribute for classifying given datasets. To determine the optimal approach for classifying a collection of learning data, it is necessary to minimize the number of questions, also known as lowering the depth of the tree. A function that can identify which questions result in the most fair and balanced distribution of replies is needed. Illustrations of such functions encompass the information gain metric. The C4.5 approach, developed by Ross Quinlan, is a tool used to generate decision trees. C4.5 is the successor to ID3. C4.5 incorporated several improvements to the ID3 algorithm. The C4.5 algorithm use the gain ratio as a statistic to choose characteristics. C4.5 categorizes attributes as either discrete or continuous. Breiman's proposed CART (Classification and Regression Tree) method is almost indistinguishable from ID3. The Gini index is used as a measure of impurity when selecting a variable in CART. Classification trees are generated for categorical target variables, whereas regression trees are created for numerical objectives with continuous values. When constructing trees, CHAID (Chi-squared Automatic Interaction Detector) employs the Chi-square contingency test in two distinct approaches. The process begins by verifying the compatibility of the predictor levels for merging. Once each predictor level has been simplified to its most significant form, the algorithm identifies the most important predictor for distinguishing between the levels of the dependent variable.

B. Bayesian Classifier. For the purpose of computing the parameters, this approach requires just a minimal amount of data preparation, which makes it not only simple but also intuitive to use. The classifier is well-structured to handle both real and different data, and irrelevant attributes are also insensitive to the classification process. A Bayesian classifier is characterized by the presence of evident class conditional independencies among its subsets of variables. Using a graphical depiction of the causal linkages between events can make the learning process easier to understand.

C. Neural Networks The immense capability of neural networks resides in their ability to extract significance from complex data for the purpose of identifying patterns and revealing potentially hazardous tendencies. An important advantage of neural networks is its capacity to uncover novel tasks through the utilization of training data or early experience. The Neural Network is capable of discovering all potential interactions between variables and predictors. The main reason for utilizing neural networks in educational data mining is to get the desired outcome [13].

D. Clustering Techniques. The process involves constructing collections of interconnected abstract entities into collections of interconnected parallel elements. Cluster analysis is conducted on a partitioned data set, where groups are generated based on data comparison. Subsequently, the following task involves allocating distinct labels to each group. Similarly, the data objects are treated as separate entities in this scenario. One of the main advantages of clustering in classification is its flexibility and use of distinct positive criteria to distinguish across groupings. Clustering techniques primarily focus on many fields, including data analysis, pattern recognition, and image processing [14].

18.4 Conclusion

Web Usage Mining is the process of using data mining techniques to find trends in online data to obtain a better knowledge of web-based services and fulfil their unique requirements. The primary goal of the Web Usage Mining approach is to evaluate data saved in log files. The registration folder of a website contains logical data related to customer needs. One possible goal of online usage mining is to make relevant recommendations to website users or to improve website organization. This study attempts to improve web caching and personalization by analyzing key online patterns, such as user groups and categories that engage with intriguing information later. Improved web cache design may be accomplished by utilizing predictive algorithms that anticipate the categories of interesting sites that will be visited next. Data about user interests gathered through group formation may be utilized to personalize the internet experience for individual visitors. This study delves thoroughly into the topics of online content mining and web use mining. In this chapter, we will discuss decision tree algorithms for online usage mining.

References

1. Chauhan, D., Singh, C., Rawat, R., Dhawan, M., Evaluating the Performance of Conversational AI Tools: A Comparative Analysis, in: *Conversational Artificial Intelligence*, pp. 385–409, 2024.
2. Bhradwaj, B., Mining educational data to analyze students performance. *IJACSA*, 2, 6, 63–69, 2011.
3. Rawat, R., Chakrawarti, R.K., Sarangi, S.K., Choudhary, R., Gadwal, A.S., Bhardwaj, V., eds. *Robotic Process Automation*. John Wiley & Sons, 2023.
4. Shahiria, A.M., Husaina, W., Rashida, N.A., A Review on Predicting Students Performance using Data Mining Techniques. *The Third Information Systems International Conference, Procedia Computer Science*, vol. 72, 2015.
5. Rawat, R., Telang, S., William, P., Kaur, U., CU, O. K. (Eds.)., *Dark Web Pattern Recognition and Crime Analysis Using Machine Intelligence*. IGI Global, 2022.
6. Pandey, U.K. and Pal, S., A Data mining view on class room teaching language. *Int. J. Comput. Sci.*, 8, 2, 277–282, 2011, ISSN:1694-0814.
7. Rawat, R., Kaur, U., Khan, S. P., Sikarwar, R., Sankaran, K. (Eds.), *Using Computational Intelligence for the Dark Web and Illicit Behavior Detection*. IGI Global, 2022, https://doi.org/10.4018/978-1-6684-6444-1.
8. Rawat, R., Telang, S., William, P., Kaur, U., C.U., O. (Eds.)., *Dark Web Pattern Recognition and Crime Analysis Using Machine Intelligence*. IGI Global, 2022, https://doi.org/10.4018/978-1-6684-3942-5.
9. Rawat, R., Chakrawarti, R.K., Sarangi, S.K., Patel, J., Bhardwaj, V., Rawat, A., Rawat, H. eds, *Quantum Computing in Cybersecurity*. John Wiley & Sons, 2023, https://onlinelibrary.wiley.com/doi/book/10.1002/9781394167401.
10. Mishra, A.K., Tyagi, A.K., Dananjayan, S., Rajavat, A., Rawat, H., Rawat, A., Revolutionizing Government Operations: The Impact of Artificial Intelligence in Public Administration, in: *Conversational Artificial Intelligence*, pp. 607–634, 2024.
11. Srivastava, M., Garg, R., Mishra, P.K., Preprocessing Techniques in Web Usage Mining: A Survey. *Int. J. Comput. Appl.*, 97, 18, 0975–8887, July 2014.
12. Namdev, A., Patni, D., Dhaliwal, B.K., Parihar, S., Telang, S., Rawat, A., Potential Threats and Ethical Risks of Quantum Computing, in: *Quantum Computing in Cybersecurity*, pp. 335–352, 2023.
13. Nahar, S., Pithawa, D., Bhardwaj, V., Rawat, R., Rawat, A., Pachlasiya, K., Quantum technology for military applications, in: *Quantum Computing in Cybersecurity*, pp. 313–334, 2023.

14. Sikarwar, R., Shakya, H.K., Kumar, A., Rawat, A., Advanced Security Solutions for Conversational AI, in: *Conversational Artificial Intelligence*, pp. 287–301, 2024.
15. Rathi, M. and Rajavat, A., High Dimensional Data Processing in Privacy Preserving Data Mining, in: *2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT)*, pp. 212–217, IEEE, 2020, April.
16. Patsariya, M. and Rajavat, A., Network Path Capability Identification and Performance analysis of Mobile Ad hoc Network, in: *2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT)*, pp. 82–87, IEEE, 2020, April.
17. Dhar, S., Dhar, U., Rajavat, A., Factors and Attributes of Team Players: A Study of Engineering Students in India, in: *International Simulation and Gaming Association Conference*, pp. 53–60, Springer International Publishing, Cham, 2021, September.
18. Nahar, S., Pithawa, D., Bhardwaj, V., Rawat, R., Rawat, A., Pachlasiya, K., Quantum Technology for Military Applications, in: *Quantum Computing in Cybersecurity*, pp. 313–334, 2023.
19. Pithawa, D., Nahar, S., Bhardwaj, V., Rawat, R., Dronawat, R., Rawat, A., Quantum Computing Technological Design Along with Its Dark Side, in: *Quantum Computing in Cybersecurity*, pp. 295–312, 2023.
20. Chauhan, D., Singh, C., Rawat, R., Chouhan, M., Conversational AI Applications in Ed-Tech Industry: An Analysis of Its Impact and Potential in Education, in: *Conversational Artificial Intelligence*, pp. 411–433, 2024.
21. Chauhan, D., Singh, C., Rawat, R., Dhawan, M., Evaluating the Performance of Conversational AI Tools: A Comparative Analysis, in: *Conversational Artificial Intelligence*, pp. 385–409, 2024.
22. Shrivastava, M., Patil, R., Bhardwaj, V., Rawat, R., Telang, S., Rawat, A., Quantum Computing and Security Aspects of Attention-Based Visual Question Answering with Long Short-Term Memory, in: *Quantum Computing in Cybersecurity*, pp. 395–412, 2023.
23. Rawat, R., Chakrawarti, R.K., Raj, A.S.A., Mani, G., Chidambarathanu, K., Bhardwaj, R., Association rule learning for threat analysis using traffic analysis and packet filtering approach. *Int. J. Inf. Technol.*, *15*, 6, 3245–3255, 2023.

19

Natural Language Processing Using Soft Computing

M. Rajkumar[1]*, Viswanathasarma Ch[2], Anandhi R. J.[3], D. Anandhasilambarasan[4], Om Prakash Yadav[5] and Joshuva Arockia Dhanraj[6]

[1]School of Computer Science Engineering and Information Systems, Vellore Institute of Technology, Tamil Nadu, India
[2]AI & DS, VIIT, Visakhapatnam, Andhra Pradesh, India
[3]Department of Information Science Engineering, New Horizon College of Engineering, Bangalore, India
[4]Department of Computer Science and Engineering, Karpagam Academy of Higher Education Coimbatore, Eachanari, Coimbatore, Tamil Nadu, India
[5]School of Computer Science and Engineering, LPU, Punjab, India
[6]Department of Computer Science and Engineering (AI&ML), School of Engineering, Dayananda Sagar University, Devarakaggalahalli, Harohalli, Kanakapura Road, Ramanagara District, Bengaluru, Karnataka, India

Abstract

The fields of text mining and information retrieval were the origins of natural language processing (NLP), which was first developed from those fields. Throughout the years, several applications that are based on artificial intelligence have developed out of its initial domain. Some examples of these applications include machine translation, query expansion, robotic command detection, and many more. The origins of natural language processing (NLP) may be traced back to a variety of disciplines, such as psychology, mathematics, computer science, linguistics, computer engineering, electrical and electronic engineering, artificial intelligence, robotics, and computer science and information. This paper provides a comprehensive analysis of a wide variety of soft computing techniques to natural language processing (NLP).

***Keywords*:** Natural language processing, soft computing, NLP challenges, support vector machine

**Corresponding author*: rajkumar.m@vit.ac.in
Joshuva Arockia Dhanraj: ORCID: 0000-0001-5048-7775

Rajesh Kumar Chakrawarti, Ranjana Sikarwar, Sanjaya Kumar Sarangi, Samson Arun Raj Albert Raj, Shweta Gupta, Krishnan Sakthidasan Sankaran and Romil Rawat (eds.) Natural Language Processing for Software Engineering, (271–282) © 2025 Scrivener Publishing LLC

19.1 Introduction

The study and development of practical applications for computers to interpret and analyze human spoken and written language is the focus of the discipline of Natural Language Processing (NLP), which includes both research and development. Researchers in the field of natural language processing (NLP) anticipate that one day they will be able to implement techniques and tools that will train computers to interpret and manipulate natural languages to carry out specified tasks [1, 2]. This will be accomplished by researching human language comprehension and usage. In the 1950s, when artificial intelligence (AI) and linguistics were just beginning to gain traction, the concept of natural language processing (NLP) started to take form. When it was first being developed, natural language processing (NLP) initially focused on text mining and information retrieval as its primary areas of interest. Throughout the years, several applications that are based on artificial intelligence have developed out of its initial domain. Some examples of these applications include machine translation, query expansion, robotic command detection, and many more. Natural language processing (NLP) is built on the foundation of a number of disciplines, including computer science, mathematics, psychology, linguistics, electrical and electronic engineering, artificial intelligence and robotics, and linguistics.

In recent years, natural language processing (NLP) has made significant progress in its development. Many of the activities that we do on a daily basis include the core applications of natural language processing. Amazon and Flipkart are two examples of online shops that make it simple to browse reviews that were made by genuine consumers. The "sentiment" that these evaluations communicate, that is, whether the product has a positive or negative review, is what decides how they are organized. Sentiment analysis, which is a job based on natural language processing, does all of this. A further excellent illustration is provided by search engines such as Google. The phrases that we type into the search engine determine whether or not it is able to offer results that are relevant to our query. In this case, natural language understanding (NLU) is the first stage, and then metadata retrieval comes next. It is vital that the search engine is able to "understand" the query that we have entered. Taking into consideration our previous searches, it ought to provide us with suggestions for relevant material whenever we key in terms such as "Match," for instance. The phrase "match" is, however, not entirely clear. This might refer to any sporting event; it is not limited to simply cricket or football. One further interpretation that may be given is the pairing process. Google will then employ query expansion to get the most accurate result possible. The use of natural language processing is also responsible for this accomplishment.

Some of the primary NLP applications are listed as follows:

- Information Retrieval
- Text Mining
- Query Expansion
- Sentiment Analysis
- Machine Translation
- Speech Recognition
- Question Answering
- Optical Character Recognition

Many have remarked that the presence of slang and polysemous terminology complicates natural language comprehension. Polysemous words may mean several things. Listeners and readers are asked to make an ongoing judgement on the appropriate interpretation depending on what is relevant to the current situation. Ambiguity is one of the most challenging aspects of natural language processing.

19.2 Related Work

Word-sense-detection (WSD) is the computational challenge of identifying the specific word-sense activation in a particular environment. The user should be able to input "That is Right," and the computer should properly determine the meaning of "right" in this context. "Correct" is one interpretation that might fit here. Human rights are another possible context here. Perhaps the "right direction" is another way of looking at it. When a term is ambiguous, its precise meaning is always deduced from its context. Word Sense Dilemma (WSD) is a well-known classic in natural language processing (NLP) that involves lexical ambiguity; several methods have been suggested to solve it [4, 5]. When thinking about ways to fix WSD, the Lesk algorithm was among the first revolutionary ideas. Words in the same "neighborhood" or "section of text" are presumed to have a similar "topic" according to the Lesk algorithm. Locating the "sense that overlaps the most between its dictionary definition and the given context" is how the Simplified Lesk algorithm determines each word's most relevant meaning. This approach takes a word-by-word approach to analyzing its meaning, rather than trying to determine the overall meaning of the text based on how other words in the same context are seen. Until recently, this algorithm was the best of its kind. In subsequent updated versions, it was further integrated. Adopted Lesk, one of the updated versions, had a 33.1% accuracy rate [5, 6].

It was tested using the SENSEVAL2 dataset. Unfortunately, this level of precision is obviously unsuitable for any real-world use. Another approach, Distributional Similarity, was suggested and explored in the literature to further enhance the accuracy. Its working hypothesis is that distributional similarity increases as the degree to which two words are "semantically similar" increases. This suggests that they are more likely to be found in comparable language circumstances. Correct contextual awareness was achieved by its utilization in several ways across different test settings [7–11]. On SENSEVAL2, this method achieved an accuracy of 39.54%. The main benefit of this method was that it introduced the concept of using semantics in WSD, which is a relatively new notion. Yet another drawback of this method was that it produced false expansions for most of the words that were evaluated.

The Translation-based Semantic Model is another method that has been published in articles. It depended on text that had been machine translated. Multiple levels of semantic analysis were used to these translations. The GOLD standard dataset was used for testing [12]. Various variants of this technique are also published in the literature [12–14]. The accuracy rates for the sports dataset are around 52.7% and the finance dataset is 58%. The main issue with this method was that it relied on translations, which weren't always accurate. An important thing to remember is that this method supported the idea of semantics in WSD [15–17].

As a result of the high levels of accuracy it achieved, WordNet graph-based approaches have grown in popularity among the aforementioned methods. Up to now, the state-of-the-art has been methods based on WordNet graphs. Numerous scholars from all around the world have delved into this area of study [17]. WordNet graphs are constructed by using a depth-first search strategy. In WordNet, the several definitions of a word serve as nodes. Semantic relations such as hypernymy, holonomy, and others form the margins. Over time, the researchers hone this technique. However, this approach has a logical flaw in that it disregards the relevance of all semantic relations. It gives each edge of the WordNet graph the same weight of 1 and assumes that all semantic relations are equal. In actuality, however, this is not the case. It appears, for example, that the hypernym "vehicle" should be given greater weight than the hyponym "sedan" when attempting to solve the "car" ambiguity. Currently available WordNet graph-based methods exclude all these factors.

Automatically constructing a summary for a given document using the main points of the original material is known as text summarization. Applying TS to a paragraph of material (10-12 lines) should result to a brief summary (3-4 lines) that provides a general overview of the full piece. All the unnecessary and superfluous details are removed. The user's reading time is preserved. Knowing the right meaning of a word is crucial for discovering the best phrases and keywords. Labeling topics is another benefit of accurate text/document summary. Generating headlines and analyzing social media and markets are two of the most common uses of text summary. Researchers have suggested Automatic Text Summarization (ATS) methods for a number of languages. As an example, the Arabic TS, or Textual Graph Model, was previously suggested in the literature and used as a standard for Arabic text summarization [18]. But this approach only managed a 28% accuracy rate when tested on 1651 papers. The lack of specification of the association factor for clustering was a drawback of this approach. Its primary value, however, is it effectively eliminated repetition from the provided text.

In addition, the LDA model was considered. The main idea was that for every paragraph of text, important subject labels could be retrieved, and then the sentences could be alphabetized. Its performance on the ROUGH2 dataset was subpar, according to the results [19]. Its main contribution, meanwhile, was elucidating the need of word frequency, phrase location, and sentence length in text sample summary. One such method that failed to handle complicated sentences was ATSSC, which used a soft computing strategy for sentence ranking [20]. Another significant method in this area that was evaluated on the ROUGH1 dataset is LexRank. For scholars, it was a watershed moment since it proposed the possibility of building text graphs based on semantics. The assessment revealed that the method's main flaw was that it just used degree centrality to choose the most important information nodes.

A large number of scholars from different parts of the world have taken an interest in studying short answer assessment. Automatic evaluation of students' brief responses is feasible. However, it is necessary to disambiguate all subject terms since students might use multiple phrases to describe the "ideal answer." Researchers have tried a number of different strategies in their quest to identify the best way to address this issue, but they have so far been unsuccessful. In [21–23], we may find descriptions of some of the major methods used in this field of study.

With optical character recognition (OCR), the goal is to make the image's text editable and machine-readable. The input might be an image file type. The picture is created by extracting the necessary text and pages from a large collection of relevant papers. OCR is useful in many aspects of daily life. After cars undergo Automated Number Plate Recognition (ANPR), traffic challans are created. Since OCR exists, this is now feasible. Using optical character recognition (OCR) software to digitize antique books and historical documents extends their storage life. It is to be expected that OCR, being a procedure that deals with images, is prone to many mistakes throughout the recording and conversion of images. Use of natural language processing methods fixes these mistakes. Disambiguation strategies are used to replace the incorrect term with the right one. Take this example: "I am going to the Aank of the river." It's rather clear that the term "Aank" is the wrong one. However, the WSD method may be required to substitute it with the proper term. Then, from the list of possibilities—"Bank," "Tank," etc.—we may choose the best term.

19.3 NLP Soft Computing Approaches

The following is a display of the typical supervised NLP machine learning algorithms that are derived from the categorized fundamental approaches

Support Vector Machine

The objective of a support vector machine, also known as an SVM, is to discover a hyperplane in a space that has N dimensions, where N is the number of features that identify the data points in a particularly distinctive manner. For the purpose of splitting the two sets of measurement data points, it is possible to investigate a number of different hyperplane concepts. Finding the best possible margin, which is defined as the biggest feasible distance between data points in both classes, is the goal of this endeavor. The ability to classify forthcoming data items is simplified by having a margin that is greater. Support vector machines, often known as SVMs, are able to classify data points based on their location on hyperplanes, which act as decision boundaries. The term "support vectors" refers to the data points that are less far from the hyperplane and have the potential to influence its positioning and orientation. The removal of these support vectors will result in a modification to the placement of the hyperplane. This is because these support vectors act as indicative points in the construction of the support vector machine (SVM), and their use will maximize the classifier margin.

Bayesian Networks

For the purpose of computing probabilities, Bayesian Networks make use of Bayesian inference, which is a kind of probabilistic graphical model. The primary purpose of this model is to illustrate, via the use of directed edges in the graph, the interdependencies and causal interactions that exist between the variables. Using these links allows for the systematic inference of factors based on the random variables that are represented by the graph. The distribution is factorized because Bayesian Networks provide a product of N components that are restricted to contingency. Naive Bayes is a well-known example that demonstrates how the most persuasive findings may be produced from the most straightforward explanations. The fact that it is simple, accurate, and consistent has allowed it to be used successfully in a variety of applications, particularly for natural language processing (NLP) challenges.

Maximum Entropy:

As a probabilistic classifier, the Max Entropy technique is related to models that use exponential functions or variables expressed as exponents. It is not presumptuous that the features must be autonomous. Using the "Principle of Maximum Entropy," the algorithm chooses the best-fitting model based on the entropy value relative to the training data. Language identification, topic classification, sentiment analysis, and other large-scale text classification problems often use the Max Entropy classifier. By modelling all known characteristics and making no assumptions about unknown ones, the maximum entropy classifier may employ dependent features to categorize texts. For each batch of training data, it takes into account all possible classifiers and chooses the one that maximizes entropy. Aspects of the text, such as unigrams and bigrams, may be mined for statistical and contextual information using this method. Classifications like "positive/neutral/negative" or "objective/subjective" are therefore used to these characteristics. Commonly used in Natural Language Processing and Knowledge Extraction is the Bag-of-Words Framework, which stores each document as a sparse array of 1s and 0s denoting the existence or absence of a particular word.

Conditional Random Fields

Conditional random fields are a kind of "Statistical Modelling Method" that sees widespread use in ML and pattern recognition. While a classifier predicts a tag for an individual image without considering "neighboring samples," a CRF may take context into consideration. It is built as a diagrammatic (graph-based) model to include interdependencies between many prognostic variables. Various kinds of graphs are used in practice, depending on their intended use. As an illustration, linear chains commonly utilized in natural language processing, CRFs are responsible for incorporating sequential dependencies into the predictions. For the purpose of developing consistent interpretations, CRFs—a form of unique directionless probabilistic graphical model—are used to discern the ensuing alliance between data. Common applications [33–37] include human word processing and other forms of linear and sequential data classification and interpretation. It is evident that CRF may be used to tasks such as point-of-sale tagging, shallow parsing, detecting entitled articles, studying genomes, and predicting important peptide portions. They are also often used in several fields where Hidden Markov Models (HMM) need to be enhanced or replaced in operation. Both "Object Recognition" and "Image Segmentation" are common applications of CRFs in computer visualization.

Artificial Neural Network

One subfield of machine learning is known as artificial neural networks (ANN), or just neural networks. To conduct the classification using ML, two sets of input data are required for the "Training" and "Testing" processes, which are derived from the way the human brain works. Neural networks are capable of identifying previously unseen patterns inside data. It is a valuable tool for data scientists that utilize Neural Network approaches to identify patterns in images, videos, audio, or text. Thanks to this approach of emotion-focused assessment, real text categorization is now within our reach. ANN can manage the relationship and correlation between input variables with the correct intricate design. Neural networks' precedence is defined by two theoretical features. Firstly, these methods are adaptable, fact-based, and capable of self-improvement. Second, every function may be correctly estimated by neural networks, thanks to a number of popular practical estimates. Typical "Recurrent Neural Networks" (RNN) were the backbone of this case. The capacity of the connected network to accept words in chronological sequence is used by vanilla RNN. Due to a simple relationship, one-dimensional Convolution Neural Networks (CNN) have outperformed RNN in terms of accuracy and may achieve a sizeable level. Summing up people's feelings on an item is what sentiment analysis is all about. As more and more technological tools become available, it is more important than ever to gauge public sentiment on goods, businesses, and personal preferences. One way to anticipate how to respond to anything on social media is to try to put oneself in the author's shoes. One may choose between social networks and online communities as their preferred kind of social media. Social networks are formed and maintained by individuals who have interacted with each other in the past; these people often seek out new connections to broaden their sphere of influence. Communities, on the other hand, consist of individuals from many walks of life and professions, with whom they often have nothing in common. One commonality among a community's members is their love of a well-known pastime.

19.4 Conclusion

It is possible to trace the origins of natural language processing (NLP) back to a wide range of disciplines, such as linguistics, computer science, mathematics, computer engineering, computer science and information, robotics, artificial intelligence, and computer and electrical engineering. The topic of natural language processing (NLP) is the primary emphasis of this article, which provides an in-depth analysis of a variety of soft computing methodologies. Support Vector Machines, Bayesian Networks, Maximum Entropy, Conditional Random Fields, and Artificial Neural Networks are some of the technologies that have shown to be useful in the processing of natural language. Within the scope of this chapter, natural language processing via the use of soft computing techniques is completely investigated.

References

1. Chowdhury, G.G., Natural language processing. *Annu. Rev. Inf. Sci. Technol.*, 37, 1, 51–89, 2003.
2. Agirre, E. and Edmonds, P. (Eds.), *Word sense disambiguation: Algorithms and applications*, vol. 33, pp. 1–12, Springer Science & Business Media, USA, 2007.
3. Namdev, A., Patni, D., Dhaliwal, B. K., Parihar, S., Telang, S., Rawat, A., Potential Threats and Ethical Risks of Quantum Computing, in: *Quantum Computing in Cybersecurity*, pp. 335–352, 2023.
4. Krovetz, R. and Croft, W.B., Lexical ambiguity and information retrieval. *ACM Trans. Inf. Syst. (TOIS)*, 10, 2, 115–141, 1992.
5. Hogaboam, T.W. and Perfetti, C.A., Lexical ambiguity and sentence comprehension. *J. Verbal Learn. Verbal Behav.*, 14, 3, 265–274, 1975.
6. Banerjee, S. and Pedersen, T., An adapted Lesk algorithm for word sense disambiguation using WordNet, in: *International conference on intelligent text processing and computational linguistics*, pp. 136–145, Springer, Berlin, Heidelberg, 2002.

7. Shrivastava, M., Patil, R., Bhardwaj, V., Rawat, R., Telang, S., Rawat, A., Quantum Computing and Security Aspects of Attention-Based Visual Question Answering with Long Short-Term Memory, in: *Quantum Computing in Cybersecurity*, pp. 395–412, 2023.
8. Noonia, A., Beg, R., Patidar, A., Bawaskar, B., Sharma, S., Rawat, H., Chatbot vs Intelligent Virtual Assistance (IVA), in: *Conversational Artificial Intelligence*, pp. 655–673, 2024.
9. Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., Soroa, A., A study on similarity and relatedness using distributional and WordNet-based approaches, in: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 19– 27, Association for Computational Linguistics, 2009.
10. Rapp, R., Word sense discovery based on sense descriptor dissimilarity, in: *Proceedings of the ninth machine translation summit*, pp. 315–322, 2003.
11. Miller, T., Biemann, C., Zesch, T., Gurevych, I., Using distributional similarity for lexical expansion in knowledge-based word sense disambiguation. *Proceedings of COLING*, vol. 2012, pp. 1781–1796, 2012.
12. Marcu, D. and Wong, D., A phrase-based, joint probability model for statistical machine translation, in: *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pp. 133–139, 2002.
13. Koehn, P., Statistical significance tests for machine translation evaluation, in: *Proceedings of the 2004 conference on empirical methods in natural language processing*, pp. 388–395, 2004.
14. Quirk, C., Brockett, C., Dolan, W., Monolingual machine translation for paraphrase generation, in: *Proceedings of the 2004 conference on empirical methods in natural language processing*, pp. 142–149, 2004.
15. Seemakurty, N., Chu, J., Von Ahn, L., Tomasic, A., Word sense disambiguation via human computation, in: *Proceedings of the acm sigkdd workshop on human*, pp. 60–63, ACM, computation, 2010.
16. Navigli, R., Faralli, S., Soroa, A., De Lacalle, O., Agirre, E., Two birds with one stone: learning semantic models for text categorization and word sense disambiguation, in: *Proceedings of the 20th ACM international conference on Information and knowledge management*, pp. 2317–2320, ACM, 2011.
17. Navigli, R. and Lapata, M., An experimental study of graph connectivity for unsupervised word sense disambiguation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32, 4, 678–692, 2009.
18. Alwan, M.A. and Onsi, H.M., A Proposed Textual Graph Based Model for Arabic Multidocument Summarization. *Int. J. Adv. Comput. Sci. Appl.*, 7, 6, 435–439, 2016.
19. Liu, N., Tang, X.J., Lu, Y., Li, M.X., Wang, H.W., Xiao, P., Topic-Sensitive Multi document Summarization Algorithm, in: *Parallel Architectures, Algorithms and Programming (PAAP), Sixth International Symposium*, pp. 69–74, IEEE, 2014.

20. Tayal, M.A., Raghuwanshi, M.M., Malik, L.G., ATSSC: Development of an approach based on soft computing for text summarization. *Comput. Speech Lang.*, 41, 214–235, 2017.
21. Suthar, H., Rawat, H., Gayathri, M., Chidambarathanu, K., Techno-Nationalism and Techno-Globalization: A Perspective from the National Security Act, in: *Quantum Computing in Cybersecurity*, pp. 137–164, 2023.
22. Bakharia, A. and Dawson, S., Using Sentence Compression to Develop Visual Analytics for Student Responses to Short Answer Questions, in: *VISLA@ LAK*, pp. 11–13, 2015.
23. Padó, U., Get Semantic With Me! The Usefulness of Different Feature Types for Short-Answer Grading, in: *COLING*, pp. 2186–2195, 2016.
24. Kouloumpis, E., Wilson, T., Moore, J., Twitter sentiment analysis: The good the bad and the omg!, in: *Fifth International AAAI conference on weblogs and social media*, pp. 1–5, 2011.
25. Pak, A. and Paroubek, P., Twitter as a corpus for sentiment analysis and opinion mining, in: *LREc*, vol. 10, no. 2010, pp. 1320–1326, 2010.
26. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R., Sentiment analysis of twitter data, in: *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pp. 30–38, 2011.
27. Saif, H., He, Y., Fernandez, M., Alani, H., Contextual semantics for sentiment analysis of Twitter. *Inf. Process. Manage.*, 52, 1, 5–19, 2016.
28. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M., Lexicon-based methods for sentiment analysis. *Comput. Ling.*, 37, 2, 267–307, 2011.
29. Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., Liu, B., Combining lexicon-based and learning-based methods for Twitter sentiment analysis. HP Laboratories, Technical Report HPL- 2011, pp. 8–19, 2011.
30. Kanjanawattana, S. and Kimura, M., Novel Ontologies-based Optical Character Recognition-error Correction Cooperating with Graph Component Extraction. *BRAIN. Broad Res. Artif. Intell. Neurosci.*, 7, 4, 69–83, 2017.
31. Vinitha, V.S. and Jawahar, C.V., Error Detection in Indic OCRs. Document Analysis Systems (DAS). *2016 12th IAPR Workshop*, IEEE, pp. 180–185, 2016.
32. Darwish, K. and Magdy, W., Error correction vs. query garbling for Arabic OCR document retrieval. *ACM Trans. Inf. Syst. (TOIS)*, 26, 1, 5–15, 2007.
33. Pithawa, D., Nahar, S., Bhardwaj, V., Rawat, R., Dronawat, R., Rawat, A., Quantum Computing Technological Design Along with Its Dark Side, in: *Quantum Computing in Cybersecurity*, pp. 295–312, 2023.
34. Shrivastava, M., Patil, R., Bhardwaj, V., Rawat, R., Telang, S., Rawat, A., Quantum Computing and Security Aspects of Attention-Based Visual Question Answering with Long Short-Term Memory, in: *Quantum Computing in Cybersecurity*, pp. 395–412, 2023.
35. Rawat, R., Chakrawarti, R.K., Raj, A.S.A., Mani, G., Chidambarathanu, K., Bhardwaj, R., Association rule learning for threat analysis using traffic

analysis and packet filtering approach. *Int. J. Inf. Technol.*, *15*, 6, 3245–3255, 2023.

36. Rathi, M. and Rajavat, A., *Analysing Cryptographic and Random Data Sanitization Techniques in Privacy Preserving Data Mining*, vol. 83, Allied Publishers, New Delhi, India, 2023.
37. Dhar, S., Dhar, U., Rajavat, A., Factors and Attributes of Team Players: A Study of Engineering Students in India, in: *International Simulation and Gaming Association Conference*, pp. 53–60, Springer International Publishing, Cham, 2021, September.

20

Sentiment Analysis Using Natural Language Processing

Brijesh Goswami[1]*, Nidhi Bhavsar[2], Soleman Awad Alzobidy[3], B. Lavanya[4], R. Udhayakumar[4] and Rajapandian K.[5]

[1]Institute of Business Management, GLA University, Mathura, UP, India
[2]Department Information Technology, Thakur College of Engineering and Technology, Mumbai, Maharashtra, India
[3]Department of English Language and Translation Studies, College of Sciences and Theoretical Studies, Saudi Electronic University, Riyadh, Kingdom of Saudi Arabia
[4]Vel Tech Rangarajan Dr Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, Tamil Nadu, India
[5]SRM Kattankulathur Dental College, SRM Nagar Kattankulathur Chengalpattu, Tamil Nadu, India

Abstract

Complex text mining techniques include text classification, topic discovery and summarization, concept extraction, document clustering, sentiment extraction, text conversion, natural language processing, and more. These techniques can then be used to extract non-trivial information from a set of text-based documents. Arguments, divergent points of view, and verbal altercations are all useful tools for presenting the facts around a current topic. Text mining, as used in natural language processing, is the process of gleaning sentiment from text that has been obtained through online networking web-based systems. This research outlines a strategy to enhance machine learning and natural language processing in an attempt to simplify and pinpoint the textual feelings that underlie information that is extracted from social media comments.

***Keywords*:** Sentiment analysis, text analysis, challenges, machine learning, preprocessing, tokenization

**Corresponding author*: brijesh.goswami@gla.ac.in
R. Udhayakumar: ORCID: 0000-0002-2447-664X

Rajesh Kumar Chakrawarti, Ranjana Sikarwar, Sanjaya Kumar Sarangi, Samson Arun Raj Albert Raj, Shweta Gupta, Krishnan Sakthidasan Sankaran and Romil Rawat (eds.) Natural Language Processing for Software Engineering, (283–294) © 2025 Scrivener Publishing LLC

20.1 Introduction

Sentiments are the feelings that a person expresses via their behavior, which influences the behavior of people around them. Sentiment analysis (SA), the study of computationally leveraging these emotions, is therefore continuing, and what counts are the outcomes. One significant natural language processing (NLP) application in this research that addresses the interplay between human and machine language is opinion mining [1]. The act of creating a system to gather and arrange data from various social media platforms, blogs, online forums, and web-based surveys is known as opinion mining. Because people base their judgements on the tools at their disposal, business establishments can gain insight into the buyer's mindset, evaluation, sentiment, and attitude through social media. Politicians and officials also make modifications while addressing public issues. Now is the moment to develop a new application using real-world data, emphasizing textual approaches for information discovery (especially on Twitter) for searching, organizing, or studying. Because Twitter is used to communicate messages via social media and blog sites, numerous unique elements in tweets also affect the different domains and methodology covered in sentiment analysis. This leads to the emergence of new difficulties [2].

Sentiment analysis provides organizations with important insights into customer sentiment, enabling them to better understand how customers respond to particular offerings. Sentiment analysis is now widely accepted by businesses, governments, and other organizations in addition to academics. Sentiment analysis is a potent technique because people's thoughts and sentiments about a certain subject can provide insightful information for a wide range of industries, including academics and business [3].

Organizations must prioritize understanding the attitudes of their customers towards their brand. Enhanced goods, in-the-moment issue detection, and distinguishing markers can all lead to happier customers. Sentiment analysis has applications across a wide range of disciplines. It gives us useful information in fields like sports, economics, politics, tourism, and healthcare and aids in our understanding of consumer behavior [4].

Sentiment analysis, a branch of natural language processing, is currently one of the most active areas in computer science study (NLP). Data extraction is made possible by genuine discourse Processing (NLP), which uses a range of machine learning algorithms to automatically examine and comprehend genuine human discourse. The authors of the survey have made recommendations for potential future improvements. They found problems with sentiment analysis and carried out a survey that could help a novice researcher tackle new tasks with remarkable accuracy. They discovered that sentiment analysis has a number of possible applications, such as predicting equity and market prices, assessing political opinion, and categorizing movie reviews to anticipate box office receipts. In conclusion, the authors feel that sentiment analysis holds great potential for further research. Sentiment analysis in social media content is a difficult undertaking because dialect articulation is diverse and varied. The difficulty of the endeavor lies not only in its complexity but also in the huge amount of real-time content available. Sentiment analysis cannot be done by hand; instead, a strong tool or intelligent system is needed [5].

20.2 Sentiment Analysis Levels

Sentiment analysis can take place at several levels, for instance, in a document or in a sentence.

20.2.1 Document Level

By employing document-level analysis, we are able to categorize the general sentiment of a reviewer after extracting it from the whole review. The main emphasis of this level is on documents that express viewpoints, whether they are positive or negative. The techniques employed for this approach yield a precision rate of 70 to 80% across several publications. This tool mostly serves the purposes of product reviews and film reviews. Sentiment analysis at the document level only considers a single entity. This strategy is unsuitable for comparing several objects or analyzing papers.

20.2.2 Sentence Level

At this stage, the initial step is to determine if a statement is subjective or objective. In this context, “personal interpretation” refers to the act of examining the statement from an individual’s distinct perspective, whereas “from a distance” it implies analyzing it from the viewpoint of another person [6]. The main objective of the second task is to determine if the subjective sentence is positive, negative, or neutral. Unlike an objective sentence, which often conveys factual information, a subjective statement expresses one’s personal views, beliefs, emotions, or values. To distinguish between the two, we utilize the Naive Bayesian Classifier technique. However, merely determining whether the sentence conveys a positive or negative attitude is inadequate. Since a subjective statement might contain subjective and factual components, as well as several views, this intermediary process is valuable for eliminating phrases that do not express any viewpoints.

20.2.3 Aspect Level

In fine-grained analysis, the prioritization of perspectives is done at the aspect level classification, which is also known as entity level, phase level, or document level. This is in contrast to language constructions such as documents, paragraphs, sentences, clauses, and phrases. By considering the positive and negative emotions expressed in a review, it enables the identification of the sentiments and characteristics associated with an object. Next, the system proceeds to detect and obtain the object attributes that have been expressed by the reviewer. The generation of the feature-based feedback summary occurs after the conclusion of synonym grouping. The principal purposes of this are to extract aspects and categorize aspect emotions [7].

20.3 Challenges in Sentiment Analysis

Ambiguity resolution, word sense disambiguation, handling out-of-vocabulary words, cultural and contextual adaptation, and domain specific challenges are some of the issues encountered at the Lexicon Level in Natural Language Processing. Addressing ambiguity necessitates mastering the nuances of words having many meanings and creating systems to accomplish accurate disambiguation in different settings at the level of the lexicon. The goal of word sense disambiguation is to resolve cases when a word might have more than one correct interpretation depending on the context. With the ever-changing nature of language and the introduction of new terms, it is essential to develop methods for dealing with words that do not appear in current dictionaries to effectively manage Out-of-Vocabulary Words. To make sure that NLP systems can understand a broad variety of expressions, it is necessary to adapt lexicons to account for cultural subtleties and contextual differences in language use. This process is called cultural and contextual adaptation. Overcoming obstacles in lexicon usage within certain domains or businesses is what “Navigating Domain-Specific Lexicon Challenges” is all about. Tailored solutions are necessary for precise processing due to the uniqueness, development, and specialization of terminology.

At the document level, Natural Language Processing has challenges such as detecting sarcasm and irony, dealing with contradictions and negations, and dealing with contextual complexity. Taking into account the larger context, such as word connections, the text's tone, and the development of emotion across a document or discussion, is an important part of contextual complexity. Understanding the many nuances of language in varied and ever-changing contexts is essential for natural language processing models when dealing with contextual complexity. Negation and contradiction are challenges that need to be overcome for sentiment analysis to be accurate in a text. These challenges might take the form of negative phrases, contradictory claims, or bad feelings. To correctly discern irony and sarcasm, one must be able to recognize and comprehend statements that reflect emotions contrary to their literal meaning. Conditional and comparative sentences, opinion word sharing, subjectivity identification and sentiment categorization, opinion source and target, handling negation, and other issues are encountered at the sentence level of natural language processing. Due to their subtle and context-dependent character, analyzing phrases that describe conditions or make comparisons requires nuanced sentiment interpretation. Identifying instances when the distribution of opinion terms among entities impacts the accuracy of sentiment analysis. When dealing with phrase level sentiment analysis, it might be difficult for natural language processors to detect subjectivity and categorize feelings. In natural language processing (NLP), knowing who expressed an opinion and where it came from, as well as how to deal with negation, are all vital parts of sentence-level sentiment analysis and opinion extraction.

Important obstacles in aspect-level sentiment analysis in NLP include aspect classification, aspect stemming, and aspect trimming. In aspect-based sentiment analysis, aspect pruning is the process of eliminating or downplaying less important or irrelevant features. Improving the analysis's efficiency and eliminating irrelevant or irrelevantly noisy data is the job of aspect trimming. Because different words might have the same root but different meanings, aspect-level analysis can be complicated due to stemming ambiguity. Operating at the aspect level presents a challenge for aspect classification in natural language processing. The challenge is in correctly classifying characteristics, making sure we comprehend subtleties in individual traits, and adjusting to the ever-changing language.

There are three tiers to these natural language processing problems: syntactic, semantic, and pragmatic. Much of the earlier work on sentiment analysis concentrated on solving problems at the syntactic layer, such as lemmatization, sentence boundary disambiguation, micro text normalization, and part of speech (POS) tagging. Problems including idea extraction, subjectivity detection, entity identification, and word meaning disambiguation are handled at the semantic layer. Practical considerations include tasks such as polarity identification, aspect level tasks, metaphor comprehension, personality recognition, and sarcasm detection.

20.4 Related Work

A number of different supervised machine learning algorithms were utilized by Pragya Juneja *et al.* (2017) to categorize the feelings that were expressed on Twitter as either positive or negative. It was the major objective of this project to make a prediction on the results of the elections for the Delhi Corporation. In the course of this work, the most effective categorization model for machine learning was also discovered. This research investigated the behavior of a number of classification models by making use of the Twitter dataset including information on various political parties. When compared to the other classifiers, the Multinomial Naive Bayes model performed much better in terms of the percentage of correct classifications. We are able to make this model even more refined to achieve even more favorable results [8].

Text analysis, natural language processing, text preparation, and stemming were some of the methods that Chhaya Chauhan *et al.* (2017) utilized in their research of the complex data. The process of deciphering human speech was accomplished via the use of a variety of computer software and methods. According to the findings of the research, the availability of natural language that is available on the internet allows for a variety of approaches to be utilized to ascertain the tone of a phrase or piece of writing. A wide range of algorithms and methods were utilized by the writers to extract [25–27] a comprehensive description of the product, item by item, to build a review that was authentic. After the system had excluded all other terms, the advice was to concentrate on higher-level natural language processing [28–30] jobs in the future. This would allow for the most efficient methods or tools to be utilized to extract more accurate results from datasets that comprised mainly keywords [9].

As a result of the growth of both persons and modern methods of communication, there has been a meteoric rise in the usage of social media websites, as stated by Jamil Hussain *et al.* (2017). It is now much simpler for individuals to convey their opinions through social media platforms such as Facebook, Blogs, Twitter, and others as a result of technological developments in CPU architectures and mobile phones. According to the findings of the study, it is possible to utilize emotion theories, machine learning techniques, and natural language processing algorithms to locate SA on a variety of social media sites. Phrase-level sentiment analysis was utilized by the authors to investigate SVM, ME, and NB, which are three of the currently available classifiers for the assessment of depression in the research. In the course of testing on Twitter and the 20 newsgroups dataset, they have included voting and feature selection algorithms. According to their findings, support vector machines (SVM) perform better than Naive Bayes in terms of accuracy [10].

Zahra Rezaei *et al*. (2017) utilized Twitter to get an understanding of the subject matter of the brief messages and the influence that they have on other people. In addition, users tweet often and swiftly. For the purpose of extracting the separation aspects of the data from Twitter, researchers utilized filtering and wrapping approaches which resulted in an improvement in SA's performance. When it came to Hoeffding trees and McDiarmid trees, they utilized the approaches that were applicable. Despite the fact that the McDiarmid tree strategy worked better than the others, SA had to work quickly to process all of the Twitter data [11].

The tweets have been further classified into seven separate emotion groups by the utilization of text-based binary and ternary classifications, as demonstrated by Mondher Bouazizi *et al*. (2017) and other organizations. To carry out the classification, they initially developed SENTA, a software that features an easy-to-use graphical user interface and provides users with the ability to select from a variety of criteria [12].

An opinion mining technique was utilized by Mondher Bouazizi *et al.* (2018) to examine Multiclass sentiment analysis when they were working on SA analysis. The purpose was not to determine the overall emotional polarity of the text message; rather, it was to determine the specific attitude of the intended recipient. They conducted an experiment using eleven distinct emotion classes, each of which was allocated to a separate internet article rather than a single tweet. The experiment was carried out utilizing the "quantification" approach and the SENTA tools. On the other hand, the data set to be used in this investigation was manually annotated, and the outcomes of the analysis were compared to those annotations [13].

Text mining and neural networks were utilized in the development of a hybrid model that was built by Mohammed H. Abd El-Jawad and colleagues (2018) to characterize sentiments. A number of deep learning and machine learning techniques were utilized in the development of the model. For this study, an analysis was performed on more than one million tweets that were collected from five distinct domains. After training with 75% of the dataset and testing with 25% of the dataset, this model exceeded the strategies that came before it with an accuracy rate of 83.7%. Additionally, the authors suggested conducting additional research, particularly in Arabic tweets, with a specific focus on the coupling of sentiment and text for the purpose of sentiment analysis. It is [14].

As stated by Mondher Bouazizi *et al.* (2019), the majority of research conducted on text SA has focused on binary and ternary data classification. The multi-class categorization of natural languages has proven to be a challenging endeavor, despite the fact that natural languages are both diverse and involved. Moreover, it is a great deal more challenging to "quantify" the manner in which individuals communicate their emotions using statistical methods. This model not only studied the difficulties associated with multi-class classification, but also discovered areas in which there is room for further development in terms of the accuracy of multi-class classification [15].

To determine whether or not a particular service or product could be appropriate for the consumer, Shahnawaz *et al.* (2017) relied on the writer's position on a particular issue. To make up for the lack of labeled data, the authors reasoned out that it would be straightforward to employ semi-supervised and unsupervised learning-based models provided that there would be sufficient unlabeled data available. This would allow them to make up for the lack of labeled data [16].

20.5 Machine Learning Techniques for Sentiment Analysis

The algorithms belong to an area of artificial intelligence called machine learning. These algorithms facilitate computer learning. Consequently, algorithms frequently acquire data sets and form assumptions on the data's quality based on their observations. This knowledge enhances the ability to forecast future data discoveries by increasing the probability of such discoveries. Given that most non-random data contain patterns that enable machines to make generalizations, it becomes very easy to forecast the hidden data [17]. A computer-trained model is employed to govern the crucial data points for the purpose of generalization. Sentiment analysis employs supervised and unsupervised machine learning methods to extract meaningful insights from both structured and unstructured textual data, with the aim of assisting decision-makers. The objective of this strategy is to tackle the distinctive challenges in sentence categorization that algorithms encounter by training a text classifier using a pre-existing dataset that has been labeled by humans. The fundamental principles of this approach are derived from both supervised and unsupervised learning [18].

Support Vector Machines:
Due to their renowned proficiency in recognizing conventional text, Support Vector Machines (SVMs) consistently outperform Naive Bayes (NB). The Support Vector Machine (SVM) is a probabilistic classifier, in contrast to Naive Bayes and MaxEnt, which prioritize accuracy. In the training phase of the two-category scenario, the primary objective is to identify a hyperplane, denoted by vector *w,* that effectively distinguishes the document vectors in each class, with a substantial separation or margin between them [18].

Random Forest (RF):
The computing system in issue is the outcome of combining several decision trees. These trees are built using an entropy metric. The purity level of a node is determined by using the value of this integer [19]. In this circumstance, classifier systems with minimal bias and significant variance are often trained since it simplifies the process of generating decision trees. There are several techniques available to enhance the efficiency of different decision trees. However, it is a frequent practice to employ the model average technique while also insuring a specific number of trees. This approach reduces the variability of the model while somewhat increasing the bias. Empirical evidence has shown that this approach enhances precision. It is crucial to randomize the construction of a feature set for each decision tree to generate diverse and impartial trees [20].

Logistic Regression:
Many individuals employ this technique to resolve binary classification problems. One of the several uses of logistic regression is in text mining, where it is used to convert tasks into binary classification. The fundamental concept underlying this approach is utilizing patterns identified in a vast dataset to forecast the level of positivity in a tweet. The model was named based on the logistic structure of the probability function [21].

Rule-based Classifiers:
A rule-based classifier employs a collection of rules to depict the data space. The left representation, denoted in standard form [22], signifies a location inside the feature set. Conversely, the right representation, denoted in class, indicates the reverse. The words are present on the page. We rarely utilize the term "absence" as it lacks significance in situations when there is

a dearth of knowledge. Each rule in the training phase is derived from one of the several rules-generation criteria that are available. The two predominant criteria are confidence and support [23]. The support training data set comprises all cases that are pertinent to the rule. Confidence refers to the conditional likelihood of the right side being satisfied, given that the left side of the rule has been met. Vader is a widely-used model that operates based on a set of rules and incorporates several lexical characteristics. Unlike conventional approaches, this method focuses on doing sentiment analysis on data gathered from microblogs and delivers comprehensive and precise outcomes [24].

20.6 Conclusion

The way people express their feelings or emotions may have a big influence on how those around them behave. Sentiment analysis (SA), the study of harnessing these emotions digitally, is therefore continuous, and the outcomes are what really count. One important NLP application in this study that addresses the relationship between machine and human language is opinion mining. The aim of text mining in the context of natural language processing is to derive meaning from text gathered via online networks and web-based systems. This work proposes an approach to enhance machine learning and natural language processing by simplifying and identifying the underlying textual sentiments of data collected from social media comments.

References

1. Anagha, M. *et al.*, Fuzzy logic based hybrid approach for sentiment analysis of Malayalam movie reviews. *2015 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)*, IEEE, 2015.
2. Shoukry, A. and Rafea, A., A hybrid approach for sentiment classification of Egyptian Dialect Tweets. *2015 First International Conference on Arabic Computational Linguistics (ACLing)*, IEEE, 2015.
3. Han, P., Li, S., Jia, Y., A topic-independent hybrid approach for sentiment analysis of chinese microblog. *2016 IEEE 17th International Conference on Information Reuse and Integration (IRI)*, IEEE, 2016.
4. Appel, O. *et al.*, A hybrid approach to sentiment analysis. *2016 IEEE Congress on Evolutionary Computation (CEC)*, IEEE, 2016.

5. Biltawi, M., Al-Naymat, G., Tedmori, S., Arabic sentiment classification: A hybrid approach. *2017 International Conference on New Trends in Computing Sciences (ICTCS)*, IEEE, 2017.
6. Jagdale, J. and Emmanuel, M., Hybrid Corrective Critic Neural Network for Sentiment Classification in Community Media. *2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA)*, pp. 1236–1241, 2019.
7. Zouzou, A. and Azami, I.E., Text sentiment analysis with CNN & GRU model using GloVe. *2021 Fifth International Conference On Intelligent Computing in Data Sciences (ICDS)*, pp. 1–5, 2021.
8. Juneja, P. and Ojha, U., Casting online votes: to predict offline results using sentiment analysis by machine learning classifiers. *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, IEEE, 2017.
9. Chauhan, C. and Sehgal, S., Sentiment analysis on product reviews. *2017 International Conference on Computing, Communication and Automation (ICCCA)*, IEEE, 2017.
10. Hassan, A.U. *et al.*, Sentiment analysis of social networking sites (SNS) data using machine learning approach for the measurement of depression. *2017 International Conference on Information and Communication Technology Convergence (ICTC)*, IEEE, 2017.
11. Rezaei, Z. and Jalali, M., Sentiment analysis on Twitter using McDiarmid tree algorithm. *2017 7th International Conference on Computer and Knowledge Engineering (ICCKE)*, IEEE, 2017.
12. Bouazizi, M. and Ohtsuki, T., A pattern-based approach for multi-class sentiment analysis in Twitter. *IEEE Access*, 5, 20617–20639, 2017.
13. Bouazizi, M. and Ohtsuki, T., Multi-class sentiment analysis in Twitter: What if classification is not the answer. *IEEE Access*, 6, 64486–64502, 2018.
14. El-Jawad, M.H.A., Hodhod, R., Omar, Y.M.K., Sentiment Analysis of Social Media Networks Using Machine Learning. *2018 14th International Computer Engineering Conference (ICENCO)*, IEEE, 2018.
15. Rathi, M. *et al.*, Sentiment Analysis of Tweets Using Machine Learning Approach. *2018 Eleventh International Conference on Contemporary Computing (IC3)*, IEEE, 2018.
16. Astya, P., Sentiment analysis: approaches and open issues. *2017 International Conference on Computing, Communication and Automation (ICCCA)*, IEEE, 2017.
17. Rabeya, T., Chakraborty, N.R., Ferdous, S., Dash, M., Al Marouf, A., Sentiment Analysis of Bangla Song Review- A Lexicon Based Backtracking Approach. *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, pp. 1–7, 2019.
18. Yang, L., Li, Y., Wang, J., Sherratt, R.S., Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning. *IEEE Access*, 8, 23522–23530, 2020.

19. Koto, F. and Rahmaningtyas, G.Y., Inset lexicon: Evaluation of a word list for Indonesian sentiment analysis in microblogs. *2017 International Conference on Asian Language Processing (IALP)*, pp. 391–394, 2017.
20. Wu, X., Linghu, Y., Wang, T., Fan, Y., Sentiment Analysis of Weak-RuleText Based on the Combination of Sentiment Lexicon and Neural Network. *2021 IEEE 6th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)*, pp. 205–209, 2021.
21. Ding, Y., Li, B., Zhao, Y., Cheng, C., Scoring tourist attractions based on sentiment lexicon. *2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, pp. 1990–1993, 2017.
22. Mehto, A. and Indras, K., Data mining through sentiment analysis: Lexicon based sentiment analysis model using aspect catalogue. *2016 Symposium on Colossal Data Analysis and Networking (CDAN)*, pp. 1–7, 2016.
23. Thavareesan, S. and Mahesan, S., Sentiment Lexicon Expansion using Word2vec and fastText for Sentiment Prediction in Tamil texts. *2020 Moratuwa Engineering Research Conference (MERCon)*, pp. 272–276, 2020.
24. Liu, J., Yan, M., Luo, J., Research on the Construction of Sentiment Lexicon Based on Chinese Microblog. *2016 8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, pp. 56–59, 2016.
25. Pithawa, D., Nahar, S., Bhardwaj, V., Rawat, R., Dronawat, R., Rawat, A., Quantum Computing Technological Design Along with Its Dark Side, in: *Quantum Computing in Cybersecurity*, pp. 295–312, 2023.
26. Shrivastava, M., Patil, R., Bhardwaj, V., Rawat, R., Telang, S., Rawat, A., Quantum Computing and Security Aspects of Attention-Based Visual Question Answering with Long Short-Term Memory, in: *Quantum Computing in Cybersecurity*, pp. 395–412, 2023.
27. Rawat, R., Chakrawarti, R.K., Raj, A.S.A., Mani, G., Chidambarathanu, K., Bhardwaj, R., Association rule learning for threat analysis using traffic analysis and packet filtering approach. *Int. J. Inf. Technol.*, *15*, 6, 3245–3255, 2023.
28. Rathi, M. and Rajavat, A., High Dimensional Data Processing in Privacy Preserving Data Mining, in: *2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT)*, pp. 212–217, IEEE, 2020, April.
29. Patsariya, M. and Rajavat, A., Network Path Capability Identification and Performance analysis of Mobile Ad hoc Network, in: *2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT)*, pp. 82–87, IEEE, 2020, April.
30. Dhar, S., Dhar, U., Rajavat, A., Factors and Attributes of Team Players: A Study of Engineering Students in India, in: *International Simulation and Gaming Association Conference*, pp. 53–60, Springer International Publishing, Cham, 2021, September.

21

Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data

C. V. Guru Rao[1]*, Nagendra Prasad Krishnam[2], Akula Rajitha[3], Anandhi R. J.[4], Atul Singla[5] and Joshuva Arockia Dhanraj[6]

[1]School of CS & AI, SR University, Warangal, Telangana, India
[2]School of Management & Commerce (SoMC), Malla Reddy University, Hyderabad, Telangana, India
[3]Institute of Aeronautical Engineering, Dundigal, Hyderabad, India
[4]Department of Information Science Engineering, New Horizon College of Engineering, Bangalore, India
[5]Lovely Professional University, Phagwara, India
[6]Department of Computer Science and Engineering (AI&ML), School of Engineering, Dayananda Sagar University, Devarakaggalahalli, Harohalli, Kanakapura Road, Ramanagara District, Bengaluru, Karnataka, India

Abstract

Web structure mining aims to identify interesting patterns within the inter-document link structure. By classifying the web pages based on the link structure's topology, the authors may infer important details about the connections and similarities between them. Associations such as those between analogous entities, entities with similar content, or entities hosted on the same server can be extracted. The structure of a domain's hyperlink hierarchy may also be found using web structure mining. This is useful for efficiently modelling the databases connected to these websites. Another aspect of online mining is web content mining. It focuses on trying to extract interesting and non-trivial patterns from the content of the web pages. A web page may include many different kinds of data, some examples of which are text, images, music, video, metadata, and links. A portion of the data is highly organized and shown as tables and HTML pages created by databases, while another portion is more semi-structured and presented as HyperText

**Corresponding author*: guru_cv_rao@hotmail.com
Joshuva Arockia Dhanraj: ORCID: 0000-0001-5048-7775

Rajesh Kumar Chakrawarti, Ranjana Sikarwar, Sanjaya Kumar Sarangi, Samson Arun Raj Albert Raj, Shweta Gupta, Krishnan Sakthidasan Sankaran and Romil Rawat (eds.) Natural Language Processing for Software Engineering, (295–310) © 2025 Scrivener Publishing LLC

Markup Language (HTML). This study project provides a thorough analysis of web data mining by examining hyperlinks, use statistics, and web content mining.

***Keywords*:** Web content mining, natural language processing, web structure mining, data mining techniques, www architecture, unstructured data

21.1 Introduction

This global computing network known as the Internet is shared by all of us. It allows all connected computers to communicate data globally. It is the cornerstone of the World Wide Web and other online services. The web is made up of all the webpages that are stored on different web servers. The amount of information and services available over the Internet is growing at an incredible pace. Every year, the number of web pages is thought to double. As the breadth and variety of the Web have increased, so too has the significance of search engines in the WWW's architecture [1].

It is a difficult effort to find and get relevant information from the Internet because of the data's extreme lack of organization. In search engines, keywords determine content. The exponential growth of WWW information sources makes it more crucial than ever for users to use automated tools to find, extract, filter, and evaluate the pertinent data and resources. Furthermore, because the Web has emerged as the primary tool for electronic commerce and because businesses have invested millions in intranet and Internet infrastructure, they also need to keep an eye on and evaluate user access habits. These factors make it imperative to create intelligent systems that, both on the client and server sides, can efficiently mine data throughout the Internet and in particular Web places [2] as shown in Figure 21.1: Web Data Mining Process.

On the internet, a number of businesses and organizations offer a wide variety of tools and applications, in addition to providing automated customer service and the ability to make purchases with the internet. The use of web-based applications and configurations for activities such as online news broadcasts, online collaboration, online education, electronic commerce, and other similar activities is rapidly becoming the general practice [3, 4]. From young children trading music files with their friends to grandparents receiving images and messages from grandchildren located all over the world, the internet is rapidly becoming an instrument that is considered to be normal for the activities that are performed by the general population on a daily basis. Even for classes that are taught in more traditional classroom settings, it is normal practice for websites that provide

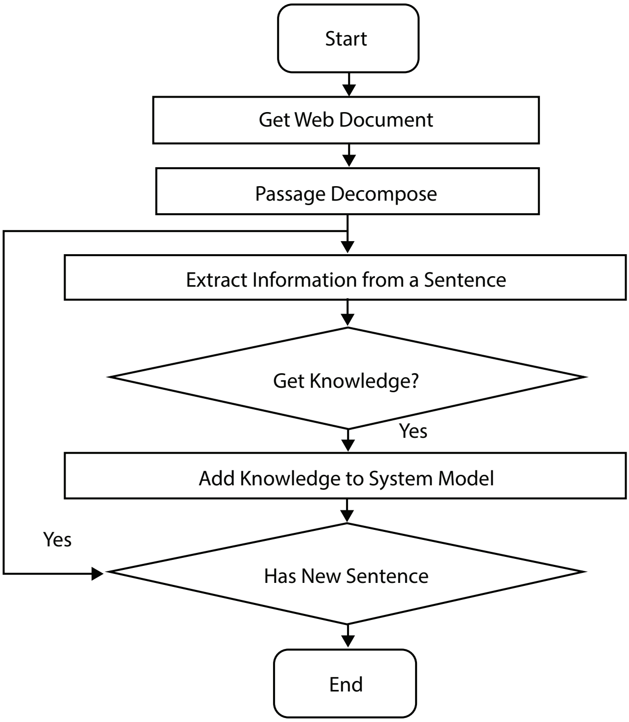


Figure 21.1 Web data mining process.

courses offered by universities and colleges to include course materials as well as extra materials. The fact that modern sophisticated distance education systems are created with the web in mind should not come as a surprise to anyone [5].

Researchers in the field of artificial intelligence, computer scientists, and information engineers who work in the field of natural language processing investigate how computers interact with human languages. More specifically, they investigate ways to teach computers to manage and comprehend large amounts of textual communication. In natural language processing (NLP), some of the most prevalent areas of difficulties are the recognition of spoken language, the comprehension of natural language, and the generation of new natural language [6]. The majority of natural language processing (NLP) systems were dependent on complex rule sets that were once constructed by hand. On the other hand, the late 1980s saw the beginning of a revolution in natural language processing brought about by the application of machine learning techniques to language processing. This development was made possible by the exponential growth of computing power as well as the diminishing influence of Chomskyan linguistic theories. The theoretical foundations of these theories have previously been a barrier to the kind of corpus linguistics that serves as the foundation for the approach that machine learning takes to language processing. In essence, the third Decision trees and other early machine learning algorithms created systems of rigorous if-then rules, which mimicked the effects of previously established rules that were already in existence. Hidden Markov models were first utilized for the purpose of part-of-speech tagging; however, statistical models have lately garnered a great deal of attention due to their capacity to make probabilistic and soft decisions by assigning real-valued weights to the features of the input data. One example of a statistical model is the cache language model, which is utilized by a significant number of voice recognition systems in the present day. These models have a tendency to be more trustworthy and more robust when given with unexpected data, particularly information that contains mistakes, when they are included into a larger system that is comprised of numerous subtasks. One of the first areas to see substantial success was machine translation, mostly as a consequence of research at IBM Research, which built more complicated statistical models. This was one of the first areas to see major development. These systems of government were able to make use of multilingual textual corpora that had been developed by the European Union and the Canadian Parliament as a result of laws that mandated the translation of all governmental proceedings into all official languages of the respective systems of government. The success of the majority of other systems, on the other hand, was and continues to be severely constrained due to the fact that these systems relied on corpora that were specifically constructed for the work that these systems intended to accomplish. The result of this is that a great deal of research has been done on techniques to learn more effectively with fewer data [7].

21.2 Web Mining

Data mining techniques are utilized in the process of web mining, which is the process of automatically locating and extracting information from documents and services that are located on the World Wide Web. This topic of study is now seeing a significant amount of growth, mostly as a result of the interest in online business. Due to the presence of this phenomenon, it is not always apparent what exactly is meant by the term “web mining” or how to compare research that pertain to this topic.

Similar to, suggest decomposing Web mining into these subtasks, namely

Data Accumulation
Gathering information from various sources, such as e-commerce websites, is the next phase. Data extraction from web documents is the primary objective. Emails, docs, newsgroups, online logs, and database transactions are all potential sources of data.

Data preprocessing
Employment and certainty are two common features of data that must be prepared. Preprocessing the collected data is essential for more efficient information mining. Accurate, succinct, and preprocessed data are necessary for data mining. Pre-processing of data includes cleaning the data, identifying users and their sessions, adding access pathways, and identifying transactions.

Pattern Discovery
Through the process of pattern identification and the use of mining algorithms, we may acquire significant and ultimately comprehensible knowledge. Several examples of such methodologies include clustering analysis, dependency modelling, classification analysis, association rule discovery, sequential pattern discovery, and so on.

Pattern Analysis
The fundamental goal of pattern analysis is to identify patterns among the patterns discovered by the model pattern discovery approach. The fundamental purpose is to identify a usable model, rules, and modes. One is capable of generating a user-friendly graphical user interface by using visualization approaches.

21.3 Taxonomy of Web Data Mining

The Web mining analysis relies on three general sets of information: previous usage patterns, degree of shared content, and inter-memory associative link structures corresponding to the three subsets in Web mining namely, as shown in Figure 21.2: Taxonomy of Web Data Mining:

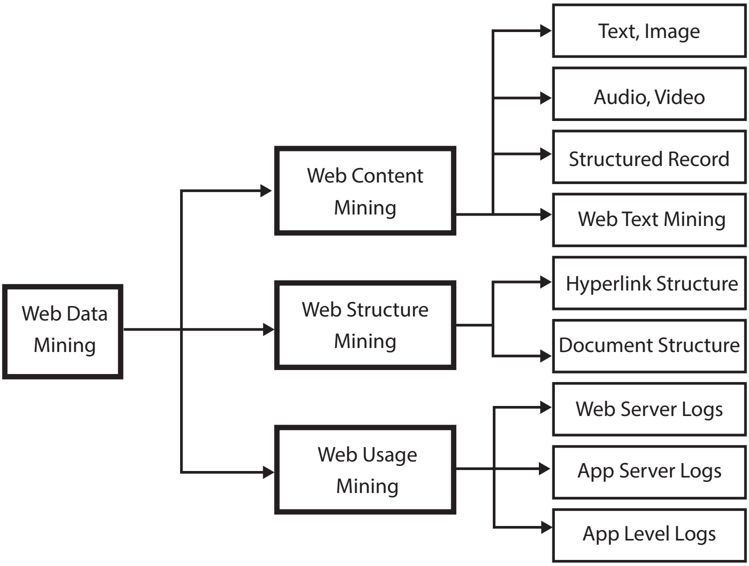


Figure 21.2 Taxonomy of web data mining.

- Web usage mining
- Web content mining
- Web structure mining

21.3.1 Web Usage Mining

The objective of web usage mining is to derive practical insights from secondary data obtained from customers' online activity. The main focus is on methods that can predict a person's behavior when visiting a website. Anticipating user behavior on the site, analyzing the differences between projected and actual usage, and customizing the site based on user preferences are all instances of mining objectives that encompass the potential strategic goals in each domain. Web use mining is same to the other two forms. Web use mining, web content mining, and web structure mining are interconnected. Clustering in pattern discovery acts as a bridge between usage mining and web content and structure mining. The web site topology will serve as the primary source of information for presenting data in web usage mining. The substantial research in the disciplines of information retrieval, databases, intelligent agents, and topology greatly enhances the benefits of web content and structure mining. Web use mining, a burgeoning topic of study, is attracting increasing attention from individuals [8].

21.3.2 Web Structure Mining

The vast majority of online information retrieval systems only use textual information, despite the fact that link data may be of great benefit. One of the goals of web structure mining is to generate a summary of the structural components that are present on the website and on each page. Web content mining is more concerned with the structure of individual documents, whereas web structure mining is more concerned with identifying the structure of the connections between documents. This is something that we can see when we break it down into its component pieces. Information such as links and similarities between websites may be generated via the use of web structure mining, which allows for the categorization of websites into different groups. Web structure mining may also result in the identification of the structure of the document, which is another potential conclusion. Using this method of structure mining, it is possible to uncover web page structures, also known as schemas. This is beneficial for navigational purposes and enables the comparison or integration of different web page schemes. Through the utilization of this structure mining technique, a reference schema will be generated, which will make it simpler to employ database techniques in order to get access to information on web pages [9].

There is a possibility that the structural information that is produced from it will be consulted for the study that is thorough on the issue. The following are elements that are included in web structure mining: the data that measures the frequency of internal and external connections in the tuples of a web table, including those that are included within the same document. This refers to the information that is used to measure the frequency of web tuples in a web table that includes both global links and linkages that span many websites. Assuming that web structure mining and web content mining are related to one another is a logical assumption to make, given that the majority of web papers have links and that both types of mining make use of main information that can be obtained on the internet. When applications are submitted, it is not unusual for these two mining occupations to be combined.

21.3.3 Web Content Mining

Data extraction from websites is the process that is known as web content mining. This technique involves automatically scanning through various online information resources. Web content mining is a branch of web mining that aims to uncover knowledge in the unstructured data that is contained inside internet pages. This is similar to the data mining methodologies that are utilized for relational databases. The most frequent types of data that may be found in online documents include text, photos, audio, video, metadata, and hyperlinks. You can also find metadata.

The majority of the data are in the form of unstructured text, although some of them are semi-structured, such as HTML pages or data in tables, or more structured, such as HTML websites created by a database. There is a need for increased complexity in web content mining due to the inherently unstructured nature of web data. The Information Retrieval View and the Database View are the two views that differentiate online content mining from other methods. For the purpose of expressing the semi-structured material, each of the works makes use of HTML structures inside the pages, and some of the works also make use of hyperlink structures between the papers. When it comes to the database viewpoint, mining is constantly trying to infer the structure of a website in order to turn it into a database. This enables better information management and querying on the web. Mining is also known as "data mining." The process of extracting useful information and insights from vast online multimedia databases is referred to as content mining. Multimedia data mining is a subset of content mining [10].

21.4 Web Content Mining Methods

Web content mining has the following approaches to mine data: Unstructured text mining, structured mining, Semi-structured text mining and Multimedia mining.

21.4.1 Unstructured Text Data Mining

The majority of the data found on the web is in the form of unstructured text. To effectively mine material, two methodologies are required: data mining and text mining. Research in the field known as text data mining, text mining, or KDT (Knowledge Discovery in Texts) focuses on applying data mining techniques to unstructured text. Some of the techniques used in text mining are

- Information Extraction
- Topic Tracking
- Summarization
- Categorization
- Clustering
- Information Visualization

21.4.2 Structured Data Mining

The Structured Data on the Web represents their host pages. Structured data is easier to extract when compared to unstructured texts. The techniques used for mining structured data are:

Web Crawler

A web crawler, often known as a spider, spiderbot, or simply crawler, is a type of Internet bot that routinely scans the whole World Wide Web, usually with the purpose of indexing it (web spidering). To keep the indexes of other sites' material up-to-date, web search engines and other websites employ web crawling or spidering software. Web crawlers save sites for further processing by search engines, which then index the pages for better user search results. Crawlers frequently access websites without permission and use system resources when they do so. When accessing huge collections of pages, issues with schedules, loads, and politeness arise. If a public website does not want to be crawled, there are ways to let the crawling agent know. For instance, by using a robots.txt file, one may instruct bots to index either certain pages or the whole website. Because there are so many sites on the Internet, not even the biggest crawlers can get them all. This is why, prior to the year 2000, search engines had a hard time providing relevant results when users entered search terms on the World Wide Web. Relevant results are provided very immediately nowadays. Humans may validate HTML code and hyperlinks using crawlers. You may use them to scrape websites as well.

Wrapper Generation

One piece of encouraging news is that the vast majority of websites that rely heavily on data make use of dynamic generation. In other words, when a user makes a request, a server script software retrieves data from a database that is located on the back end of the system and then fills it into a certain HTML template. In most cases, the pages in question include a large number of instances of data objects that adhere to a particular alignment and format. This particular kind of template is described by the wrapper.

21.4.3 Semi-Structured Data Mining

Semi-structured data has emerged as a solution to enable the accurate representation of complex real-world entities without imposing unnecessary constraints on application developers. This evolution has moved away from rigidly ordered relational databases that primarily store strings and numbers. HTML is the only language that has such a distinctive intra-document structure. The techniques utilized in semi-structured data mining encompass

21.5 Efficient Algorithms for Web Data Extraction

A genetic algorithm might be used to automatically eliminate the information that is deemed to be bad from a number of different websites. In the event that any extraneous components have been eliminated, the data are subsequently transmitted to the trinity structure.

Trinity is a method for gathering content from the internet in a process that is automated. Due to this particular reason, having connection to the internet is essential. The three basic methods that it uses to sort the data that it has gotten from the web are the prefix, the separator, and the suffix. Furthermore, it computes the estimation problem that the system is experiencing. Immediately following the retrieval of the material from a website, the stemming process is utilized to do an automated cleaning of the material. A technique known as "Ant Colony Optimization" is utilized in order to extract the content that is relevant from the webpage. Ant colony optimization produces accurate results when it is done correctly. An absence of the NP-Completeness property may be observed in it. There is a beneficial framework for data that are provided by its application. The authors' methodology allows for the detection of additional online contents that have been retrieved.

During the process of partial alignment, only the fields in a collection of data records that are capable of being aligned or matched with absolute confidence are aligned. All of the other fields in the data set are disregarded by it. It searches the internet for records of data and then extracts data from those records. The data records are separated by means of visual information, which allows for the identification of each record that is contained on a page. Following that, it inserts the data into a table that is stored in the database. With this method, the data records are initially gathered from a web page in order to facilitate the alignment process. To extract data, only records of data that have been officially acknowledged are utilized. Because there are k trees involved in the process, the complexity of the approach is O (K2).

This method, which is based on Hadoop MapReduce, is employed for the purpose of integrating, extracting, and traversing web pages on a massive scale. The procedures of navigation, extraction, and integration that are involved in the process of obtaining data from the web are all carried out without interruption. The data record set is compiled from a significantly large number of online articles that are similar to one another through the use of a loop arrangement. Through the use of the Hadoop Map Reduce framework, the three phases of the extraction model as well as the rules for regulating the loop are parallelized. The strategy of "divide and conquer" should be utilized. The implementation of a rule holder that is referred to as a page model will result in each and every web page having its own individual set of rules for the extraction of data and the navigation of the internet. Creating the Mapper class is done with the intention of making the execution of parallel extraction easier. A specific search page will serve as the beginning point for its navigation, retrieval, and integration of results.

21.6 Machine Learning Based Web Content Extraction Methods

In their 2017 study, Qingtang Liu and colleagues proposed an automated method for extracting the core content of web pages. As a means of identifying the characteristics of web page attributes, this method makes use of text density and hyperlink density as markers and indicators. An estimating approach may detect whether a node is content or noise by using the attributes of the node and its neighbors. This is done in accordance with a continuous distribution of page content. Using this strategy, we are able to filter out navigation that is irrelevant as well as ad nodes. The results of an experiment conducted on ten different news websites showed that this algorithm was successful in achieving an acceptance rate of 96.34% on average [11].

According to Vijay and Prasadh's description, a framework was developed for the processing of deep web pages in multi-data regions. The framework makes use of an enhanced co-citation approach that extracts the visual information of deep web pages directly from the web database. This is done rather than developing a separate set of application programming interfaces (APIs) [7, 15–17] for the purpose of visual information extraction. Empirical experiments with a broad range of databases demonstrate that the suggested vision-based technique (VBEC) may give high accuracy, enabling efficient and accurate recall value of comparable queries with improved time consumption, and outperform other approaches. This is demonstrated by comparing the VBEC to alternative extraction procedures [12].

The authors Narendra *et al.* proposed a method for the aggregation of information that involves the extraction of certain portions of the contents of web pages. The process of extracting paragraphs from web pages [18–22] is a work in progress. For this aim, one should make use of a parsing system that is capable of deciphering HTML pages and URL files. Java queries are those that are used as our data extraction and manipulation tools. To accomplish the task of gathering fragmented content from a number of different websites, a recommendation was made. The jsoup library is required to extract paragraphs from an HTML text that is not currently being accessed online. During the process of obtaining web pages, the approach that has been presented takes into consideration the queries that users have entered, and then separates those pages into partial pieces for each individual user [13].

The two-stage Map-Reduce strategy to data warehouse optimization that is known as Chabok was created by Barkhordari and Niamanesh [14]. In this method, the intermediate results are transmitted to the Reducer once the aggregation process has been completed on the Mappers. Omission may be joined by Chabok without the need to replicate data. For the purpose of benchmarking, the solution that was described was implemented on Hadoop, and TPC-DS queries were executed. In terms of the amount of time it takes to execute queries, Chabok fared better than other prominent big data solutions for data warehousing.

Mythili and Vetriselvi [7] constructed a DOM tree for a web content and utilized a Genetic algorithm (GA) to calculate fitness values in order to identify the optimal number of clusters in a dataset. This was done to find the optimal quantity of clusters. We have broadened the scope of our study to incorporate the WCOLT technique for classification and the COATES algorithm for clustering. Both of these methods are founded on the content of the web as well as auxiliary components. Because of the incorporation of these techniques into web mining, noisy blocks are eliminated, which results in significantly improved outcomes for online data mining. To train a machine learning model, Wu *et al.* [15] generated many features by making use of the properties of the nodes that make up a DOM tree. Afterwards, the learning model is utilized to ascertain the nodes that are contenders for the position. To choose possible nodes that are lacking data, one needs to develop a clustering method. After that, noisy data should be filtered out based on the idea that true content is frequently located in a geographically continuous block. By doing exhaustive testing on an actual dataset, you can demonstrate that the solution generates high-quality results and outperforms a number of baseline methodologies.

Nguyen-Hoang *et al*. proposed the Genre-Oriented Content Extraction (GOCE) framework as a novel approach to the process of content extraction. The first stage is the categorization of online genres, and the second stage is the development of algorithms for content extraction based on the detected genre. To assess the genre of a website based on an image that is presented, GOCE begins by employing cutting-edge models that are derived from convolutional neural networks. Following that, the content extraction algorithm is utilized in conjunction with the genre information to enhance the process of textual content extraction. GOCE has been shown to function effectively in trials, as seen by the results of the end-to-end content extraction task and the online genre classification [16].

21.7 Conclusion

One kind of data mining is known as online content mining, and its primary objective is to identify relevant patterns within the information posted on websites. A web page may contain a number of different kinds of data, including text, photographs, audio, video, metadata, and links, among other sorts of data. There are some data that are extremely well organized and presented in tables created by the database as well as HTML pages. On the other hand, there are some data that are more semi-structured and displayed in Hyper Text Markup Language (HTML). The purpose of this study project is to provide a complete analysis of web data mining by analyzing web content mining, use statistics, and connections.

References

1. Mishra, A.K., Tyagi, A.K., Dananjayan, S., Rajavat, A., Rawat, H., Rawat, A., Revolutionizing Government Operations: The Impact of Artificial Intelligence in Public Administration, in: *Conversational Artificial Intelligence*, pp. 607–634, 2024.
2. Al-asadi, T.A. and Obaid, A.J., Discovering similar user navigation behavior in web log data. *Int. J. Appl. Eng. Res.*, 1, 16, 8797–8805, 2016.
3. Munshi, A. and Tanna, S., An Improved Approach to Find Frequent Web Access Patterns from Web Logs. *Int. J. Adv. Res. Innov. Ideas Educ.*, 2, 3, 1183–1190, 2016.

4. Chitraa, V. and Thanamani, A.S., Clustering of navigation patterns using Bolzwano_Weierstrass theorem. *Indian J. Sci. Technol.*, 8, 12, 1–9, 2015.
5. Garcin, F., Dimitrakakis, C., Faltings, B., Personalized news recommendation with context trees. *Proceedings of the 7th ACM Conference on Recommender Systems*, pp. 105–112, 2013.
6. Hussain, T., Asghar, S., Masood, N., Web usage mining: A survey on preprocessing of web log file. *2010 International Conference on Information and Emerging Technologies (ICIET)*, pp. 1–6, 2010.
7. Mythili, S. and Vetriselvi, T., Analytics of Noisy Data in Web Documents Using a Dom Tree. *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, 5, 4, 1473–1480, 2015.
8. Nadee, W. and Prutsachainimmit, K., Towards data extraction of dynamic content from JavaScript Web applications, in: *2018 International Conference on Information Networking (ICOIN)*, pp. 750–754, IEEE, 2018.
9. Nagappan, V.K. and Elango, P., Agent-based weighted page ranking algorithm for Web content information retrieval. *International Conference on Computing and Communications Technologies (ICCCT)*, pp. 31–36, IEEE, 2015.
10. Nesi, P., Pantaleo, G. *et al.*, A Distributed Framework for NLP-Based Keyword and Keyphrase Extraction from Web Pages and Documents, in: *Proc. of 21st Int. Conf. on Distributed Multimedia Systems (DMS2015)*, 2015.
11. Liu, Q., Shao, M. *et al.*, Main Content Extraction from Web Pages Based on Node Characteristics. *J. Comput. Sci. Eng.*, 11, 2, 39–48, 2017.
12. Vijay, R. and Prasadh, K., A Vision-Based Approach for Web Data Extraction Using Vision-Based Approach for Web Data Extraction Using Enhanced Cocitation Algorithm. *Int. J. Comput. Sci. Issues*, 10, 5, 2, 1694–0784, 2013, ISSN (Print): 1694-0814 | ISSN (Online).
13. Jathe, N.M., Nayana, K., Netware B., *et al*, Comparative Analysis of Various Methodology to Detect Paragraph from Web Document. *Int. J. Eng. Dev. Res.*, 5, 2, 1437–1440, 2017.
14. Barkhordari, M. and Niamanesh, M., Chabok: a Map-Reduce based method to solve data warehouse problems. *J. Big Data*, 5, 1, 40, 2018.
15. Rathore, N. and Rajavat, A., Smart Farming Based on IOT-Edge Computing: Applying Machine Learning Models For Disease And Irrigation Water Requirement Prediction In Potato Crop Using Containerized Microservices, in: *Precision Agriculture for Sustainability*, pp. 399–424, Apple Academic Press, USA, 2024.
16. Patsariya, M. and Rajavat, A., A Progressive Design of MANET Security Protocol for Reliable and Secure Communication. *Int. J. Intell. Syst. Appl. Eng.*, *12*, 9s, 190–204, 2024.
17. Rathi, M. and Rajavat, A., Investigations and Design of Privacy-Preserving Data Mining Technique for Secure Data Publishing. *Int. J. Intell. Syst. Appl. Eng.*, *11*, 9s, 351–367, 2023.

18. Dubey, P. and Rajavat, A., Effective K-means clustering algorithm for efficient data mining, in: *2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN)*, pp. 1–6, IEEE, May, 2023.
19. Chauhan, D., Singh, C., Rawat, R., Chouhan, M., Conversational AI Applications in Ed-Tech Industry: An Analysis of Its Impact and Potential in Education, in: *Conversational Artificial Intelligence*, pp. 411–433, 2024.
20. Chauhan, D., Singh, C., Rawat, R., Dhawan, M., Evaluating the Performance of Conversational AI Tools: A Comparative Analysis, in: *Conversational Artificial Intelligence*, pp. 385–409, 2024.
21. Wu, S., Liu, J. *et al.*, Automatic web content extraction by a combination of learning and grouping, in: *Proceedings of the 24th international conference on World Wide Web, International World Wide Web Conferences Steering Committee*, pp. 1264–1274, 2015.
22. Nguyen-Hoang, B.D., Pham-Hong, B.T. *et al.*, Genre-Oriented Web Content Extraction with Deep Convolutional Neural Networks and Statistical Methods, in: *Proceedings of the 32nd Pacific Asia Conference on Language, Information, and Computation*, 2018.

22

Intelligent Pattern Discovery Using Web Data Mining

Vidyapati Jha[1]*, Chinnem Rama Mohan[2], T. Sampath Kumar[3], Anandhi R.J.[4], Bhimasen Moharana[5] and P. Pavankumar[6]

[1]*Department of Computer Applications, NIT Raipur, Chhattisgarh, India*
[2]*Department of Computer Science and Engineering, Narayana Engineering College, Nellore, Andhra Pradesh, India*
[3]*School of CS & AI, SR University, Warangal, Vijay, Telangana, India*
[4]*Department of Information Science Engineering, New Horizon College of Engineering, Bangalore, India*
[5]*School of Computer Science and Engineering, Lovely Professional University, Punjab, India*
[6]*Vardhaman College of Engineering, Hyderabad, India*

Abstract

Finding trends in people's internet activity and subsequently customizing their experiences is the aim of web usage mining. In online usage mining, the functional information derived from the data was employed. It gathers information from web log records to comprehend how users interact with websites. There are several approachable research projects and helpful resources available for certain purposes. The soon-to-be-collected data can be used for customization, structure enhancement, website modification, and industrial intelligence and usage characterization. A technique known as "web usage mining" is used to collect information from real users about how they navigate websites. Web usage mining, also known as web log mining, aims to identify intriguing and frequently occurring user access patterns from web browsing data stored in web server logs, proxy server logs, or browser logs. Personalization, system improvements, corporate intelligence, advertising, and design enhancement are just a few of the numerous applications for web usage mining. The goal of this research is to find intelligent patterns in web usage mining.

**Corresponding author*: vjha.phd2021.mca@nitrr.ac.in

Rajesh Kumar Chakrawarti, Ranjana Sikarwar, Sanjaya Kumar Sarangi, Samson Arun Raj Albert Raj, Shweta Gupta, Krishnan Sakthidasan Sankaran and Romil Rawat (eds.) Natural Language Processing for Software Engineering, (311–322) © 2025 Scrivener Publishing LLC

***Keywords*:** Pattern discovery, web usage mining, machine learning, graph base methods

22.1 Introduction

A branch of data mining called "web mining" focuses on how far the field has advanced in terms of retrieving valuable information from the internet. Web mining focuses mostly on techniques for extracting insightful information from data that is stored on the internet. Data mining, the science of analyzing data from several perspectives to find novel and fascinating patterns, trends, correlations, and patterns of useful knowledge, is applied to web data. This thesis describes the study of using publicly accessible benchmark data to find interesting online patterns, such as web page navigation sequences, through the application of graph traversal and data mining techniques. An overview of web mining techniques is given at the beginning of this chapter, followed by a description of the datasets used and their application to online page navigation sequences. The structure of the remainder of the thesis is outlined in the section that follows. Driven by the exponential rise in internet users, a large number of researchers are exploring the web domain in an attempt to find interesting and non-trivial patterns. By identifying useful and engaging patterns, they may enhance the online user experience and draw in new visitors while maintaining existing ones. The technique of finding important information connected to various online pages is known as web mining. Web mining mostly targets websites that provide electronic services. "E-services" refers to a broad category of online activities, such as banking, government, advertising, education, and shopping [1].

Discovering interesting patterns in the inter-document link structure is the aim of web structure mining. By classifying web pages based on the structure of their link structures, important details about the connections and commonalities among them can be inferred. It is possible to infer relationships that are synonymous, have similar content, are hosted on the same server, and have other similar affiliations. Web structure mining can also be used to learn the details of a domain's link hierarchy. This is helpful in mimicking the functionality of the databases that these websites employ.

Another aspect of online mining is web content mining. Its objective is to extract meaningful and non-trivial patterns from the text of online pages. A web page can host various kinds of data including text, images, audio, videos, metadata, and hyperlinks. A fraction of this data is more formally arranged in tables and HTML pages produced by databases, and the remainder is provided in a semi-structured HTML style. On the other hand, most of the content on the websites are disorganized. The immense challenges presented by the need to manage such different data sources, in addition to unstructured data, call for sophisticated and intricate approaches [2].

Web mining also includes the analysis of user behavior on the internet. Here, the primary focus is on data mining techniques to discover useful patterns from secondary data that are gathered from customers' online activity. As a user browses the internet, an audit trail that lists every website they have visited is maintained. Furthermore, a substantial amount of other system-related data is also retained. The web servers log every request a user makes for a web page, no matter where they come from.

22.2 Pattern Discovery from Web Server Logs

An analysis of the web server logs uncovers fascinating patterns and a plethora of valuable insights into the web pages and user behavior. Analysis and processing of log entries can reveal valuable insights into highly desirable web pages, captivating websites, categories of intriguing web pages accessed thereafter, communities of users sharing common interests, identification of both fraudulent and legitimate sessions, patterns in web traffic behavior, consumer preferences in purchasing, and numerous other findings. Uncovering these fascinating patterns helps improve the online environment by personalizing it, optimizing web caching, boosting systems, making site changes, and providing business intelligence, among other methods.

Personalization refers to the practice of customizing a user's online experience based on their previous internet activities. Web cache optimization involves implementing modifications to web caches to minimize user latency to the greatest extent possible. To enhance the system, it is necessary to study the patterns of web traffic to determine appropriate rules for network transmission. This, in turn, enhances the quality of service in terms of speed and other performance measures. Site modification refers to the process of redesigning websites based on feedback from users and insights gained from studies on user behavior on the web. Business intelligence aims to improve the online environment by focusing on the increasing number of e-marketing sales. The patterns in which people travel through online pages are highly valuable data that may be collected from web server logs and utilized in the indicated fields. The identification of each user's sessions is achieved by preprocessing the data obtained from the web server logs [3].

Using the user ID in combination with the timestamp allows for determining the order of sessions, at a minimum. The scenario dictates the pre-established length of a session. The user's subsequent page visits during that session can be determined by aggregating all of the page requests made by the user inside that one-hour time frame, assuming the time duration is set to that specific period of time. One can acquire extensive knowledge regarding user groups, both fraudulent and genuine sessions, web traffic patterns, consumer purchasing preferences, following web page visits, highly interconnected and loosely interconnected web pages, as well as trends within these navigation sequences, among other things [4].

22.2.1 Subsequently Accessed Interesting Page Categories

In the future, intriguing page categories that online users are interested in can be found by analyzing and processing web page navigation sequences. Knowing which web page categories are most often requested, which ones aren't, and which ones ought to be given priority while building a web cache is made easier with the use of this data. Web cache has a limited amount of storage; thus, it should be used wisely by storing the most interesting content first. The removal of these pages ought to be a secondary priority. On the other hand, the least crucial documents should be thrown first when you're cleaning up the storage [5].

22.2.2 Subsequent Probable Page of Visit

The subsequent probable page of visit pattern provides information about the next page that the visitor may view based on his previous browsing behavior. This pattern is useful in a variety of applications, including web page recommendation systems, pre-fetching of potential visit pages, configurable page design, and others.

22.2.3 Strongly and Weakly Linked Web Pages

Reorganizing the website to draw in more visitors can be aided by the information offered by pages with strong and weak links. Web pages that are heavily connected or frequently browsed can be easily reached with the help of quick links. Conversely, poorly linked web pages can be improved or eliminated to increase their popularity within the website.

22.2.4 User Groups

Online communities that are comprised of individuals who share a common interest are extremely important from a variety of perspectives. The concept of "similar interest" is also multi-faceted in its own right. According to the general rule, people who use the internet have a tendency to congregate together based on the interests that they have in common. One further method for categorizing website users according to their interests is to scrutinize the manner in which they browse around a website. If one user browses websites about recopies and cleaning the house, while another user browses websites about kitchen appliance maintenance and space management, then the user's interest in home arranging and maintenance might be said to be shared. It is possible to improve this customer's online experience by advertising products and services such as house cleaning supplies, storage organizers, and online food delivery services, among other things [6].

22.2.5 Fraudulent and Genuine Sessions

A person may tell if a session was started by a real user or a fake one by looking at their click streams and the order in which they navigated the web pages. By limiting the ability to continue, detecting this pattern can help prevent many fraudulent online activities.

22.2.6 Web Traffic Behavior

Among the numerous useful insights gleaned from studying web traffic patterns are the peak request times of day, the times of day with relatively few requests, the average amount of time users spend online, and much more. It is also necessary to frame policies based on traffic to preserve customer service quality due to the deluge of data transfer and the fast growth of video on demand services.

22.2.7 Purchase Preference of Customers

Aligning with customer interests to increase company profitability is the goal of the pattern linked with consumer purchasing preferences. This pattern's principal goal is to supplement the website's business intelligence with the purpose of improving its marketing strategies. Based on their browsing habits and previous purchases, customers can be shown a wide variety of relevant products. Our objective is to ensure client retention and satisfaction by providing exceptional service. The customer lifetime value, the effectiveness of promotional activities, cross-marketing strategies, policies on targeted offers and coupons, and many other important pieces of information can be derived from this pattern [7, 8].

To extract these patterns from the navigation sequences of web pages, efficient procedures are something that is required. Providing assistance to the community of online researchers would be an excellent example of social service, and methodological approaches that are based on data mining and graph traversal would be ideal for this circumstance. An overview of the numerous approaches that may be utilized to extract knowledge patterns from web server logs is shown in the next section.

22.3 Data Mining Techniques for Web Server Log Analysis

The extraction process requires a variety of techniques for each pattern to be successful. A summary of the core extraction approach is presented here in few sentences. The three primary aspects of web usage mining are data pre-processing, pattern discovery, and pattern analysis. There are three key components. The navigation sequences of web pages are subjected to pre-processing before being uploaded into data mining algorithms for the purpose of identifying patterns [9].

A number of studies have demonstrated that data mining techniques are employed with the purpose of uncovering useful information that is concealed within the data that is provided. Both supervised and unsupervised learning models are the two primary categories of data mining learning models.

Association rule mining and categorization are two examples of models that can be used for supervised learning. For the purpose of classification, models are formed by making use of training data, for which the class labels are already known. Using these models, we are able to make an educated guess as to the category of data whose label was previously unknown. Decision trees are among the most frequent choices for displaying the categorization rules that are used as a result of the classification process. Because of this, analysis and interpretation of the data that are provided are made easier [10].

There are three different kinds of nodes that can be found in a decision tree: root nodes, internal nodes, and leaf nodes. When a decision tree is constructed, the internal nodes are used to represent the characteristics.

The branches are used to display the possible values of these attributes, and the leaf nodes are used to indicate the ultimate option or classification label. Typically, the root node and its recursive subdivisions are the building blocks from which the complete tree is constructed in a tree structure. To construct the decision tree, the attribute values of the data provided are utilized. The first step in the process of creating a decision tree is to identify the particular characteristic that is readily apparent in distinguishing the various scenarios [11].

The information gain, gain ratio, gain index, and several other assessment metrics are utilized to ascertain which attribute would be capable of performing the most effective discriminating performance. In the process of constructing the tree, we make use of the attribute that has the greatest or lowest assessment metric, depending on the metric that we select. If all the data instances in a particular category have the same class value for the attribute that was chosen, then the branch is terminated, and the class that was obtained is applied to the leaf node. In situations when this does not work, we select an additional characteristic that corresponds to the assessment measure that is either the best or the worst. This is something that we continue to do until either we discover the ideal mix of traits for a certain class or we exhaust all the attributes that we have. When we are still unable to arrive at a choice that is unambiguous based on the information that we have, we assign the branch to the class that has the greatest number of instances under it.

It is not necessary for unsupervised data mining systems to have the ground truth of training data for them to learn information. Researchers state that clustering is one of the most widely used unsupervised approaches among the many strategies. A similarity measure is used to divide the data into groups, and the objective is to achieve a low intra-cluster distance while simultaneously achieving a high inter-cluster distance. Clustering algorithms handle the number of groups in a variety of different ways; some of them decide on it while the grouping process is taking place, while others receive it as an input and maintain it at a constant value [12].

Association rule mining is a technique that can be applied in either a supervised or unsupervised manner, depending on whether or not a particular purpose is present. Association rule mining is a technique that is used in data mining. This technique presents an opportunity to discover interesting associations between the attributes of a dataset. During the process of rule evolution, a metric of interest is utilized. To begin, the learning model will generate candidates. After that, it will prune to locate groups of objects that are encountered frequently. During the process of pruning, it is a common practice to use the minimum support that must be addressed as a guidance. If somebody wants to proceed with the process of developing association rules, the next step is to make sure that the frequent item sets meet the minimum confidence level. To generate candidates, it is necessary to create each and every conceivable combination of the components. The count and support of an object are computed to locate collections of objects that are frequently encountered. Following that, association rules are developed by making use of the frequent item sets that satisfy the minimum confidence requirement. In the following step, candidates of the next greater size are generated by using the combinations that were left out and that satisfy the interestingness criterion. In this instance, candidates are eliminated when their levels of support and confidence are insufficient. The procedure is restarted whenever there are no combinations that satisfy the interestingness metrics [13].

As a result, the association rules are constructed using antecedents and consequents as their fundamental components. The antecedents are the sections that are on the left side of the sentence, and the consequents are the elements that are on the right side of the sentence. By doing an analysis [10–13] of the antecedents and the consequents either separately or in conjunction with one another, it is feasible to arrive at intriguing findings. Jaccard similarity coefficient is a method for determining the degree to which the groups are comparable to one another. Estimating the degree to which different groups are similar is the principal application of this tool. To put it succinctly, it is an intersection over the matter of union process. Comparing the sizes of the two sets or groups that are being investigated and then dividing that number by the cardinality of the intersection is the method that is used to determine the degree of similarity. The Jaccard similarity coefficient between any two groups has to be as low as is practically possible, taking into consideration how dissimilar the groups ought to be from each other. Through the use of pattern analysis, one can have a better understanding of the outcomes of the data mining method of pattern identification, which in turn assists in the process of drawing conclusions.

22.4 Graph Theory Techniques for Analysis of Web Server Logs

Pattern analysis, pattern discovery, and data pre-processing are all components of graph theory approaches to web server log data analysis [14–17]. Graph theory also includes pattern analysis. The pre-processing in this context includes a number of different tasks, including the cleaning of data, the identification of users and sessions, the creation of navigation sequences for web sites, and the visualization [14, 15] of these tasks in graphs. Following that, methods that are founded on graph theory are utilized to discover patterns. Following that, the patterns that were obtained are analyzed to ascertain the degree of interest the patterns possess [18, 19].

Before anything else, the sequences of the web page are subjected to the standard preparation methods. Next, a similarity calculation is performed to determine whether or not it is possible to group pages that are similar to be together, whether or not it can assist with trimming, and so on. It is possible to use either a weighted or binary adjacency matrix to represent the patterns of internet access or the attributes that are obtained from them. Because of this, it is possible to create graphs that are either directed or undirected. Through the process of traversing this graph, users who share a great deal of commonalities are subsequently grouped together.

The process of traversing a graph involves going to each and every node in the graph (West 2001). Depending on the sequence in which people are visited, graph traversal algorithms are typically utilized in two different ways and are commonly used. Breadth First Search (BFS) and Depth First Search (DFS) are two examples of traversal algorithms. In the process of conducting a depth-first search, the child nodes are visited before the sibling nodes. To put it in another way, the depth of the graph is what is utilized for exploration, not the breadth of the graph. Stack is the data structure that is utilized to put this traversal strategy into action. However, while utilizing breadth, the sister nodes are explored first rather than the child nodes. This is because the child nodes are not traversed. A data structure known as a queue is utilized to make this traverse function properly. After beginning with a queue or stack that is empty and continuing until the graph contains no more nodes that have not been visited, user groups can be constructed by making use of the nodes that are already present in the graph. Following the collection of the user groups, we proceed to investigate each one in greater depth to discover fascinating insights.

22.5 Conclusion

Online use mining gathers information from online log records to better understand how users behave on websites. One can select from a variety of doable research assignments and helpful resources, depending on his/her requirements. The soon-to-be-acquired data can be used for customizing, structuring better, modifying websites, and obtaining intelligence and industrial usage insights. Researchers can gain additional insight into the "web usage mining" behaviors of real people by watching how they interact with the internet. The aim of web usage mining, also known as web log mining, is to look for intriguing and frequently occurring user access patterns by searching through web server logs, proxy server logs, or browser logs. Applications for web usage mining are numerous and include advertising, corporate intelligence, system improvements, customization, and design optimization. Our goal in this investigation is to identify astute patterns in online usage mining.

References

1. Vakali, A., Pokorny, J., Dalamagas, T., An Overview of Web Data Clustering Practices, 2013.
2. Ananthi, J., A Survey Web Content Mining Methods and Applications for Information Extraction from Online Shopping Sites. *Int. J. Comput. Sci. Inf. Technol.*, 2014.
3. Singh, S. and Kaur, S., Web Log File Data Clustering Using K-Means and Decision Tree. *Int. J. Adv. Res. Comput. Sci. Software Eng.*, August 2013.
4. Rana, H. and Patel, M., A Study of Web Log Analysis Using Clustering Techniques. *Int. J. Innov. Res. Comput. Commun. Eng.*, June 2013.
5. Mehtaa, P., Parekh, B., Modi, K., Solanki, P., Web Personalization Using Web Mining: Concept and Research Issue. *Int. J. Inf. Educ. Technol.*, October 2012.
6. Suguna, R. and Sharmila, D., clustering web-log files-A review. *Int. J. Eng. Res. Technol.*, April 2013.
7. Singh, S. and Badhe, V., An Exclusive Survey on Web Usage Mining For User Identification. *Int. J. Innov. Res. Comput. Commun. Eng.*, 2014.
8. Adhvaryu, R. and Vidyapith, G., A Review Paper on Web Usage Mining and Pattern Discovery. *J. Inf. Knowl. Res. Comput. Eng.*, Nov 12 to Oct 13.
9. Aldekhail, M., Application and Significance of Web Usage Mining in the 21st Century: A Literature Review. *Int. J. Comput. Theory Eng.*, February 2016.
10. Patel, R. and Kansara, A., Web Usage Mining- A survey on User's Navigation pattern from Web Logs. *Int. J. Sci. Res. Dev.*, 2014.
11. Chirgaiya, S. and Rajavat, A., Tiny object detection model based on competitive multi-layer neural network (TOD-CMLNN). *Intell. Syst. Appl.*, 18, September 2022, 200217, 2023.
12. Rathi, M. and Rajavat, A., *Analysing Cryptographic and Random Data Sanitization Techniques in Privacy Preserving Data Mining*, vol. 83, Allied Publishers, New Delhi, India, 2023.

13. Dhar, S., Dhar, U., Rajavat, A., Factors and Attributes of Team Players: A Study of Engineering Students in India, in: *International Simulation and Gaming Association Conference*, 2021, September, Springer International Publishing, Cham, pp. 53–60.
14. Nahar, S., Pithawa, D., Bhardwaj, V., Rawat, R., Rawat, A., Pachlasiya, K., Quantum Technology for Military Applications, in: *Quantum Computing in Cybersecurity*, pp. 313–334, 2023.
15. Pithawa, D., Nahar, S., Bhardwaj, V., Rawat, R., Dronawat, R., Rawat, A., Quantum Computing Technological Design Along with Its Dark Side, in: *Quantum Computing in Cybersecurity*, pp. 295–312, 2023.
16. Chauhan, D., Singh, C., Rawat, R., Chouhan, M., Conversational AI Applications in Ed-Tech Industry: An Analysis of Its Impact and Potential in Education, in: *Conversational Artificial Intelligence*, pp. 411–433, 2024.
17. Chauhan, D., Singh, C., Rawat, R., Dhawan, M., Evaluating the Performance of Conversational AI Tools: A Comparative Analysis, in: *Conversational Artificial Intelligence*, pp. 385–409, 2024.
18. Shrivastava, M., Patil, R., Bhardwaj, V., Rawat, R., Telang, S., Rawat, A., Quantum Computing and Security Aspects of Attention-Based Visual Question Answering with Long Short-Term Memory, in: *Quantum Computing in Cybersecurity*, pp. 395–412, 2023.
19. Rawat, R., Chakrawarti, R.K., Raj, A.S.A., Mani, G., Chidambarathanu, K., Bhardwaj, R., Association rule learning for threat analysis using traffic analysis and packet filtering approach. *Int. J. Inf. Technol.*, *15*, 6, 3245–3255, 2023.

23

A Review of Security Features in Prominent Cloud Service Providers

Abhishek Mishra[1]*, Abhishek Sharma[2], Rajat Bhardwaj[3], Romil Rawat[4], T.M.Thiyagu[5] and Hitesh Rawat[6]

[1]STA, SVIIT, SVVV Indore, Indore, Madhya Pradesh, India
[2]SVIIT, SVVV Indore, Indore, Madhya Pradesh, India
[3]Department of Computer Science and Engineering, ASET, Amity University, Bengaluru, KA, India
[4]Department of CSE, Shri Vaishnav Vidyapeeth Vishwavidyalaya, Indore, India
[5]Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, India
[6]Department of Business Management and Economics, University of Extremadura, Badajoz, Spain

Abstract

Today, cloud storage has become an important part of the computing model because it provides resources that are cost-effective, flexible, and less likely to fail. On the other hand, safeguarding issues in the broader adoption industry face serious challenges. The diversity of the cloud and the multitude of places where data is stored further exacerbate these problems. The main points discussed about protection are; availability, recognition, integrity, authenticity and confidentiality. Encouraging more customers to move to the cloud will allow service providers to deliver secure data. Electronic forensic software, encryption methods and intrusion detection mechanisms are some of the guarantees of a reliable service in recovering and collecting evidence of intrusion activities. This article outlines some of the issues, pitfalls, and solutions associated with cloud platforms. It also includes the use of major cloud service providers for data storage and protection.

***Keywords*:** Cloud security, AWS, GCS, oracle, alibaba, sales force, IBM

**Corresponding author*: Abhishekmishra559@gmail.com

Rajesh Kumar Chakrawarti, Ranjana Sikarwar, Sanjaya Kumar Sarangi, Samson Arun Raj Albert Raj, Shweta Gupta, Krishnan Sakthidasan Sankaran and Romil Rawat (eds.) Natural Language Processing for Software Engineering, (323–334) © 2025 Scrivener Publishing LLC

23.1 Introduction

The term cloud is derived from the diagram where the Internet is represented as a cloud symbol [1]. Cloud Computing (CC) is a network that optimizes resources and services at lower cost [2]. CC is a standard that enables unified, simple, on demand network access to a shared set of computing resources (such as networks, servers, storage, usage, and services) accessible to a minimum of Management and/or administrator. The National Institute of Standards and Technology (NIST) [3] states that cloud computing is a shift in technology that uses the internet. Virtual Data Center (VDC) is efficient and provides the hardware, software and data needed to create CC services. The goal of the CC concept is to enable users to get all the work done without having to purchase hardware or software. Organizations using CC eliminate infrastructure costs by paying for the resources they use. Tools for privacy, access control, monitoring, encryption, and other capabilities are available on Amazon Web Services (AWS). It is known for its openness and flexibility in its joint security role. Confidentiality, data integrity and hardware security are the three key areas that Google Cloud Provider (GCP) focuses on. Azure gives the same priority to security management for hybrid cloud deployments [4]. Some security vulnerabilities in cloud computing still remain a major problem in the cloud environment. CC is growing as more and more companies embrace the use of the cloud, but it also brings with it many security features. Many companies are reluctant to entrust sensitive data to the cloud due to the various risks involved. The use of virtual machines brings with it many problems in terms of security and privacy of cloud services. Sending information over the Internet is another problem. Multiple users can share cloud resources using multiple strategies. Creating a secure architecture is also under the influence of this concept. Cloud service providers refuse to include access control or security monitoring and protection measures due to obvious problems. To protect important data and users from external or internal threats and ensure availability, integrity and confidentiality, some cloud service providers offer secure operation. There are always some security issues with major cloud service providers as shown in Table 23.1: Attack Breaches on cloud providers.

23.2 Cloud Computing Overview

CCs can be divided into different groups, for example, by deployment and service. These classifica-tions help understand the differences and concepts of cloud computing. The main categories of service examples are as follows:

Table 23.1 Attack breaches on cloud providers.

Cloud service provider	Security breach	Year
AWS	Data from 1 billion users was exposed by a misconfigured S3 bucket.	2019
IBM	The user's financial information is shared by the cloud server without security.	2019
Salesforce	Phishing attempts to obtain customer information by using user cre-dentials.	2018
Alibaba Cloud (ABC)	Marketing Cloud misconfiguration allowed unauthorized access.	2019
GCP	Nicknamed "Gooligan" - Malware exploited vulnerabilities on Android devices to steal Google account credentials. Not a direct GCP breach, but exposed access to GCP resources if leveraged.	2016
Azure	Azure Cosmos DB misconfiguration accidentally exposed customer support data (2.5 million records).	2022

Infrastructure as a Service (Iaas): Provide users with the ability to install and run software, including operation and use, as well as operation, storage, networking and other important financial services. The operating system, storage, installed applications, and possibly limited control of certain network components are entirely under the customer's control.
Platform as a Service (PaaS): Give users options to purchase or upload home inventory live directly to the property using approved providers, programming languages, and tools. The applications used and the configuration of the application hosting environment will be under the control of the customer; However, the networks, servers, operating systems and storages that make up the cloud infrastructure are not controlled and managed.
Software as a Service (Saas): allows users to access and use applications created by cloud service Providers. Web browsers and other user interfaces allow users to access applications from a variety of devices, such as webbased email. Except for some specific users choosing the application configuration option, customers have no control over the network, servers, operating systems, storage, or even the functionality of an application in the cloud architecture. For cloud architecture solutions, deployment methods have been discovered and are explained below:

Public cloud: Cloud infrastructure is owned by a company that provides cloud services and is publicly traded or available to large business groups.
Private Cloud: Cloud infrastructure provides more control, privacy, and customization because it is devoted to a single company.
Hybrid Cloud: integration of public and private clouds, enabling the sharing of apps and data between them.
Community Cloud: shared by multiple groups with related issues
Multi Cloud: This includes improving resilience, avoiding vendor lock-in, and utilizing services across several cloud providers.

23.3 Cloud Computing Model

CC Framework. Based on their involvement, as depicted in Figure 23.1, there are five main players in CC [3]. One who uses services from a cloud provider and pays for them according to usage is known as a cloud consumer, or cloud service consumer (CSC). Cloud services are supplied to the CSC by a cloud provider, often known as a cloud service provider (CSP). An unbiased evaluation of the performance, security, and information system operations of cloud deployments is carried out by a cloud auditor. To facilitate commercial transactions, the cloud broker works as a liaison between CSP and CSC. It is the cloud carrier who connects CSP and CSC to the cloud and offers cloud services. The cloud can be called private cloud, public cloud, community cloud and hybrid cloud. Our delivery models IaaS, PaaS and SaaS are widely accepted and legal as shown in Figure 23.1 CC Actors.

A collection of technologies, guidelines, regulations, and practices known as cloud security are used to safeguard cloud-based infrastructure, data, and

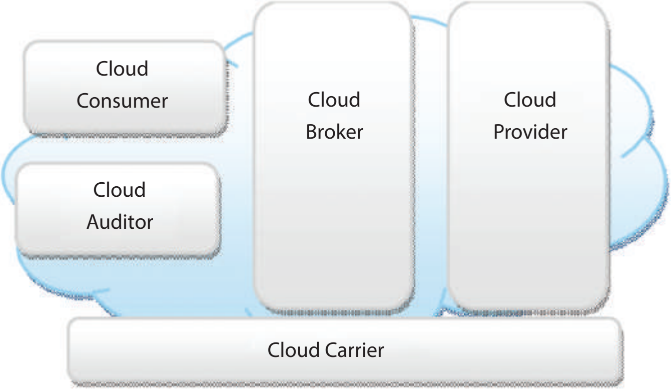


Figure 23.1 CC Actors.

Table 23.2 Features provided by service model.

Component	SaaS	PaaS	IaaS
Application Protection	CSP	CSC	CSC
Platform Protection	CSP	CSP	CSC
Services	CSP	CSC	CSP
Endpoint Protection	CSC	CSC	CSC
Data Protection	CSC	CSC	CSC
Network Protection	CSP	CSP	CSC
User Protection	CSC	CSC	CSC
Containers Workloads	CSC	CSC	CSC
APIs and Middleware	CSP	CSC	CSC
Code	CSC	CSC	CSC
Virtualization	CSP	CSP	CSC

systems. Ensuring the confidentiality, integrity, and availability of resources housed in the cloud requires a multi-layered approach. Both the consumer and the cloud service provider (CSP) share responsibility for cloud security, although their roles are different as shown in Table 23.2: Features provided by Service Model. The functions of CSP and CSC in the service delivery models are clearly defined in the Table 23.3- Various security issues in cloud.

23.4 Challenges with Cloud Security and Potential Solutions

Any system can experience security issues due to attacks, malfunctions, misconfigurations, vulnerabilities, or physical vulnerabilities. Many aspects of CC contribute to cloud security issues. An important aspect of computing today is cloud security, which deals with the protection of processes, data and applications in the cloud environment. While the cloud has many benefits, including cost-effectiveness, scalability, and flexibility, it also presents unique challenges and security issues that companies must address to protect their assets. Various issues in CC security and their solutions are listed in Table 23.3 below.

Table 23.3 Various security issues in cloud.

Cloud security problem	Description	Challenges	Mitigation strategies
Data Breaches and Loss of Data Control	Unauthorized access or leakage of sensitive data.	Lack of control over data storage and processing in the cloud, shared responsibility model.	Encrypt sensitive data, implement acces controls, conduct regular audits, and educate users on security best practic
Insecure Interfaces and APIs	Vulnerabilities in cloud service interfaces and APIs.	Lack of standardization, poor implementation, and inadequate security controls.	Update your API regularly, use secure coding practices, and use good authentication and authorization processes.
Inadequate Identity, Credential, and Access Management	Insufficient control over user identities, credentials, and access.	Weak authentication, compromised credentials, and improper access controls.	Implement robust identity and access management (IAM), enforce strong authentication, and regularly review a revoke unnecessary permissions.
Insecure Configuration of Cloud Services	Improperly configured cloud resources leading to vulnerabilities.	Complexity of cloud configurations, lack of expertise, and misconfigurations.	Regularly audit and monitor configurati employ automation for security compliance, and follow security best practices for each cloud service.
Insider Threats	Security risks posed by employees, contractors, or business partners.	Malicious or unintentional actions by authorized users.	Implement least privilege access, conduc regular employee training, monitor u activities, and enforce strong insider threat detection measures.
Compliance and Legal Issues	Violation of regulatory requirements and legal obligations.	Varying regulatory landscapes, lack of transparency, and evolving compliance requirements.	Stay informed about regulatory requirements, conduct regular compliance assessments, and work w legal and compliance experts.

(*Conti*

Table 23.3 Various security issues in cloud. (*Continued*)

Cloud security problem	Description	Challenges	Mitigation strategies
Data Loss Prevention (DLP)	Measures to prevent unauthorized access, sharing, or theft of sensitive data.	Ensuring data privacy and compliance, dealing with different data types, and preventing accidental data loss.	Use encryption, classify and label sensit data, implement robust DLP solution and monitor data access and transfer
Multi-tenancy Risks	Shared infrastructure and resources among multiple users or tenants.	Isolation concerns, resource contention, and potential for cross-tenant attacks.	Implement strong isolation mechanisms conduct regular security assessments and choose cloud providers with rob multi-tenancy security measures.
Denial of Service (DoS) and Distributed Denial of Service (DDoS)	Disruption of services by overwhelming resources.	High-profile targets, resource limitations, and evolving attack methods.	Implement DDoS mitigation solutions, Content Delivery Networks (CDNs), and design scalable architectures wit redundancy.
Lack of Transparency and Visibility	Limited insight into the security controls and processes implemented by the cloud provider.	Uncertainty about security practices, monitoring capabilities, and incident response procedures.	Choose cloud providers with transparen security practices, use third-party security tools, and conduct regular audits and assessments.

Table 23.4 Various Security measures used by major cloud providers.

Security aspect	**AWS**	**Azure**	**GCP [8]**	**Alibaba cloud [7]**	**Sales force**	**IBM cloud**
Identity and Access Management (IAM)	Robust IAM with fine-grained access control	Azure Active Directory for identity management	Cloud IAM for access control	Resource Access Management (RAM)	Salesforce Identity	IBM Cloud IAM
Network Security	Virtual Private Cloud (VPC), Network ACLs	Virtual Network (VNet), Network Security Groups (NSGs)	VPC, Firewall Rules	VPC, Security Groups	Salesforce Shield	VPC, Security Grou
Data Encryption	Encryption at rest and in transit	Azure Disk Encryption, SSL/TLS for data in transit	Encryption at rest and in transit	Data Encryption Service	Platform Encryption	Encryption at rest ar in transit
Compliance and Certifications	ISO 27001, SOC 2, HIPAA, GDPR, etc.	ISO 27001, SOC 2, HIPAA, GDPR, etc.	ISO 27001, SOC 2, HIPAA, GDPR, etc.	ISO 27001, SOC 2, HIPAA, GDPR, etc.	ISO 27001, SOC 2, HIPAA, GDPR, etc.	ISO 27001, SOC 2, HIPAA, GDPR,
Security Monitoring and Logging	AWS CloudWatch, AWS Config, AWS CloudTrail	Azure Monitor, Azure Security Center, Azure Log Analytics	Cloud Monitoring, Cloud Audit Logging	CloudMonitor, Log Service	Event Monitoring, Salesforce Shield	IBM Cloud Monitor and IBM Cloud Activity Tracker
Distributed Denial of Service (DDoS) Protection	AWS Shield	Azure DDoS Protection	Cloud Armor	Anti-DDoS	Not specified	DDoS Protection
Incident Response	AWS Incident Response, AWS WAF	Azure Security Center	Google Cloud Security Command Center	Not specified	Incident Response	Not specified

(*Contin*

Table 23.4 Various Security measures used by major cloud providers. (*Continued*)

Security aspect	**AWS**	**Azure**	**GCP** [8]	**Alibaba cloud** [7]	**Sales force**	**IBM cloud**
Container Security	AWS ECR, AWS Fargate, AWS App Mesh	Azure Kubernetes Service (AKS), Azure Container Registry (ACR)	Google Kubernetes Engine (GKE), Container Registry	Container Service, Security Center	Salesforce Functions, Heroku	IBM Cloud Kuberne Service (IKS), Container Regis
Serverless Security	AWS Lambda, AWS Step Functions	Azure Functions, Azure Logic Apps	Cloud Functions, Cloud Run	Function Compute, Serverless Workflow	Salesforce Functions	IBM Cloud Functio IBM Cloud Cod Engine
Threat Detection and Prevention	Amazon Guard Duty	Azure Security Center	Cloud Security Scanner	Alibaba Cloud Threat Detection Service	Not specified	IBM Cloud Security Advisor

23.5 Comparative Analysis

The cloud security [9–13] includes threats, security issues and possible attack preventions [5]. The shift towards zero trust architecture is an important trend in cloud security, driven by the need for stronger access controls and more effective protection against advanced threats [6, 12–15]. The comparative analysis of several aspects provided by different service providers that completely satisfy customers' high demand for cloud services [16, 17] is displayed as below in Table 23.4: Various Security measures used by major cloud providers.

23.6 Conclusion

CC has advantages such as fast use, money saving, large storage capacity and easy access to the system anytime, anywhere. It strives to support users by providing a seamless and complete experience, regardless of resources. Cloud computing appears to be a very fast technology and global computing environment. The main issues regarding CC are data security and privacy. Businesses that are aware of security threats and attacks are rapidly using the cloud. This comparative analysis with the help of forums will help users understand and choose better services according to their needs and help them get better social decisions, scalability, implementation, connectivity, distribution and support. Since the CC is a very fast changing engine, many new features have been added. The framework needs to be improved by adding more features but for now the comparison is based on the features and technologies provided by the open source here.

References

1. Zissis, D. and Lekkas, D., Addressing Cloud Computing Security Issues. *Future Gener. Comput. Syst.*, 28, 3, 583–592, 2012.
2. Ali, M., Khan, S.U., Vasilakos, A.V., Security in cloud computing: Opportunities and challenges. *Inf. Sci.*, *305*, 357–383, 2015.
3. Hogan, M. and Sokol, A., NIST Cloud Computing Standards Roadmap Version 2. NIST Cloud Computing Standards Roadmap Working Group, NIST Special Publications 500-291, NIST, Gaithersburg, MD, pp. 1–113, 2013.

4. Guptha, A., Murali, H., Subbulakshmi, T., *Proceedings of the Fifth International Conference on Intelligent Computing and Control Systems (ICICCS 2021)*, 2021, ISBN: 978-0-7381-1327-2.
5. Mishra, A., Gupta, N., Gupta, B.B., Security Threats and Recent Countermeasures in Cloud Computing, in: *Modern Principles Practices and Algorithms for Cloud Security*, pp. 145–161, IGI Global, USA, 2020.
6. Olaoye, G. and Luz, A., Future trends and emerging technologies in cloud security, 2024.
7. Luqman, A., Mahesh, R., Chattopadhyay, A., Privacy and Security Implications of Cloud-Based AI Services: A Survey. arXiv preprint arXiv:2402.00896, 2024.
8. Manthiramoorthy, C. and Khan, K.M.S., Comparing several encrypted cloud storage platforms. *Int. J. Math. Stat. Comput. Sci.*, *2*, 44–62, 2024.
9. Chirgaiya, S. and Rajavat, A., Tiny object detection model based on competitive multi-layer neural network (TOD-CMLNN). *Intell. Syst. Appl.*, 18, September 2022, 200217, 2023.
10. Rathi, M. and Rajavat, A., *Analysing Cryptographic and Random Data Sanitization Techniques in Privacy Preserving Data Mining*, vol. 83, Allied Publishers, New Delhi, India, 2023.
11. Dhar, S., Dhar, U., Rajavat, A., Factors and Attributes of Team Players: A Study of Engineering Students in India, in: *International Simulation and Gaming Association Conference*, 2021, September, Springer International Publishing, Cham, pp. 53–60.
12. Nahar, S., Pithawa, D., Bhardwaj, V., Rawat, R., Rawat, A., Pachlasiya, K., Quantum Technology for Military Applications, in: *Quantum Computing in Cybersecurity*, pp. 313–334, 2023.
13. Pithawa, D., Nahar, S., Bhardwaj, V., Rawat, R., Dronawat, R., Rawat, A., Quantum Computing Technological Design Along with Its Dark Side, in: *Quantum Computing in Cybersecurity*, pp. 295–312, 2023.
14. Chauhan, D., Singh, C., Rawat, R., Chouhan, M., Conversational AI Applications in Ed-Tech Industry: An Analysis of Its Impact and Potential in Education, in: *Conversational Artificial Intelligence*, pp. 411–433, 2024.
15. Chauhan, D., Singh, C., Rawat, R., Dhawan, M., Evaluating the Performance of Conversational AI Tools: A Comparative Analysis, in: *Conversational Artificial Intelligence*, pp. 385–409, 2024.
16. Shrivastava, M., Patil, R., Bhardwaj, V., Rawat, R., Telang, S., Rawat, A., Quantum Computing and Security Aspects of Attention-Based Visual Question Answering with Long Short-Term Memory, in: *Quantum Computing in Cybersecurity*, pp. 395–412, 2023.
17. Rawat, R., Chakrawarti, R.K., Raj, A.S.A., Mani, G., Chidambarathanu, K., Bhardwaj, R., Association rule learning for threat analysis using traffic analysis and packet filtering approach. *Int. J. Inf. Technol.*, *15*, 6, 3245–3255, 2023.

24

Prioritization of Security Vulnerabilities under Cloud Infrastructure Using AHP

Abhishek Sharma[1]* and Umesh Kumar Singh[2]

[1]Shri Vaishnav Vidyapeeth Vishwavidyalaya, Indore, India
[2]Institute of Computer Sciences, Vikram University, Ujjain, India

Abstract

In an ideal world, cyber security and IT experts would aggressively find and patch every possible vulnerability, ensuring that their enterprises would be safe across all known attack vectors. Conventional vulnerability management solutions just aren't up to the task in this fast-paced world. Despite this, cybersecurity and IT professionals are always under pressure to keep the business secure from the avalanche of vulnerabilities. The CVSS system is used by most businesses to prioritize vulnerability control operations. It's a free, open-source methodology for determining the severity of vulnerabilities. CVSS, on the other hand, fails as a prioritizing metric because it lacks the specificity to discern between the highest levels of severity — how can anything be "critical" if everything is? The majority of businesses understand the importance of having a priority strategy. In the absence of any strategy, the organizations will be faced with an overwhelming volume of effort and will have to make near-random decisions on what to mitigate first. Hence, to mitigate all the vulnerabilities in the system is practically not feasible. So, there is a strong research scope for a framework or model which is capable of prioritizing the security vulnerability which will be further helpful for the mitigation. The objective here is to propose novel security attributes matrix which can be used as input for proposed method. A novel approach and model is introduced for prioritization of security vulnerabilities for Cloud computing environments using Analytical Hierarchical Process (AHP).

Keywords**:** Cloud computing, vulnerabilities assessment, cloud security, prioritization, mitigation plan, Machine Learning Model (MLM), Artificial Intelligence (AI), Analytical Hierarchical Process (AHP)

**Corresponding author*: abhiujn9@gmail.com

Rajesh Kumar Chakrawarti, Ranjana Sikarwar, Sanjaya Kumar Sarangi, Samson Arun Raj Albert Raj, Shweta Gupta, Krishnan Sakthidasan Sankaran and Romil Rawat (eds.) Natural Language Processing for Software Engineering, (335–356) © 2025 Scrivener Publishing LLC

24.1 Introduction

The delivery of IT resources is becoming dominated by the business model known as "cloud computing," which is an evolution of information technology. With cloud computing, people and businesses may access managed, scalable IT resources like servers, storage, and apps on demand [1]. The majority of cloud users erroneously believe that cloud providers would handle "entirely" cloud security. That kind of complacency is flawed. Certain aspects of the cloud (referred to as "cloud security") are within the control of cloud providers, while other aspects (referred to as "cloud security") are the responsibility of cloud users. According to reports, Cloud Providers were not technically at fault in the great majority of data breaches; rather, the breaches were caused by security or access misconfigurations that Cloud Users had made. The [2], describes the shared responsibility matrix between Cloud Providers and various types of Cloud Users. By modelling the suggested SRM-based Cloud Computing security framework in their own or other cloud systems, cloud actors may assess security features and compliance. The author of [3] examined the worries and difficulties with the cloud computing architecture at several levels, such as application, network, host, and data. Maintaining the security of cloud computing is important, especially for the infrastructure. Several research works have been conducted in the cloud infrastructure security area; however, some gaps have not been completely addressed, while new challenges continue to arise. However, digital transformation has created new economic potential, it has also generated additional risk areas. The corporate cyber-attack surface now includes cloud infrastructures, mobile devices, DevOps, IoTs/IIoTs and key assets.

In [4], the author describes how to build an experimental setup using open-source cloud service provider. Open stack platform is one of them which has to be evaluated through the study. To investigate the security issues following open-source platforms are used for building a cloud based experimental setup, the list of OSS are as follows. It is required to set up some application over the cloud experimental setup like LMS, CRM, CMS, etc. The authors also describe how they used the cloud computing architecture to model and build an online learning solution. It focuses on how all stakeholders may be included in the teaching-learning process and how to build and implement a fully managed learning management system (LMS) using open-source on a cloud computing platform. There are numerous security and protection concerns as a result of the ongoing growth in the number of established Cloud Service Providers (CSPs).

In [5], the author discusses the security purposes and difficulties of Cyber Physical Systems, analyses the security risks and assaults on three layers of Cyber Physical Systems, and describes these issues. A wide range of security flaws, from network-level to application-level, can affect cloud technologies. Cloud service security outlines the many types of current cloud assaults as well as the successfully implemented measures to reduce the risks.

In [6], the author looks into some of the significant security problems and difficulties associated with the cloud computing environment. The reasons behind the vulnerabilities that are exceptional to the cloud are discussed, which helps with security risk analysis and identification. In the future, risk analysis and evaluation could be carried out using vulnerabilities to safeguard cloud data from hackers. The authors of [7] discuss the significant security threats and difficulties connected to cloud, which are existing in this study. Additionally, causes of cloud-specific vulnerabilities are discussed, which facilitates identification & evaluation of safety threats. Also provided is a taxonomy of protection-related dangers and challenges. Authors in [8] established that cloud environments, including virtualization and SOA, are to blame for various security flaws. Additionally, the author compares and investigates the main security problems and difficulties between in-house datacenter-based deployment and public cloud computing environments like AWS and GCP.

The authors' main goal in [9] is to create an experimental model to reduce characteristics, achieve the best accuracy on the test data set, and identify the best existing schemes used to the dataset. The behavior of different machine-learning algorithms is examined to show a danger factor related to cloud computing. By arbitrarily dividing the sub dataset into four different groups, the influence of the testing and training subsets of the information is described. The results of the experiment show that partitioning the dataset into 95% - 5% yields the highest yield of all the remaining partitions. They also show that the Decision Tree Classifier algorithm, which is used in cloud computing, performs better than randomizable filter classifier across all data sets.

To solve these problems, the author of [10] proposes an artificial intelligence-enabled access control mechanism (AI-ACM) including vehicle nodes and roadside units (RSUs). For V2I deployment, roadside units (RSUs), which are regularly fixed, are necessary to gather information on the current state of local traffic. As a result, to raise automotive environmental consciousness, V2I requires a wide coverage of RSUs.

A concerning 312% rise, or 20173 new Common Vulnerabilities and Exposures (CVEs), were published in 2021 compared to 6454 in 2016 [11]. 16,500 new CVEs were discovered in 2018. Patching every vulnerability is just not possible because the typical organization discovers 870 CVEs across 960 IT assets every single day (according to the report published in tenable). The problem set must be scaled down by organizations. The Common Vulnerability Scoring System (CVSS) is used by many businesses to rank what needs to be fixed. However, CVSS cannot significantly improve operational effectiveness on its own. Instead, companies must distinguish between vulnerabilities that pose hypothetical risks and those that actually pose risks, and then, they must order those vulnerabilities according to the level of risk they pose.

24.2 Related Work

According to Gartner (June 4–7, 2018), "Through 2021, reducing vulnerabilities will be the single most important enterprise effort to improve security." But vulnerability mitigation is more difficult than ever. In addition to traditional IT-managed assets, the cyberattack surface has grown to include cloud, DevOps, mobile, and web infrastructure, as well as newly linked hardware like IP-enabled operational technology and IoT devices. The issue of vulnerability overload is made worse by this increase in assets. To prioritize software vulnerabilities, the authors in [12] utilized a hybrid Fuzzy best-worst approach (FBWM) that took decision-makers' complexity and subjectivity into account. The author also discovered that the most serious vulnerabilities, Information Gain (IG), SQL Injection (SQLI), and Gain of Privileges, need to be fixed as soon as they are feasible. However, it is only applicable to software flaws like SQL injection, etc. In [13], the author's primary goal in developing the cost model is to determine the ideal discovery and patch release time so that the overall developer cost can be reduced while taking risk considerations into account. A high exploitability reported vulnerability in Google Chrome has been explored at the source level to demonstrate the suggested methodology.

In [14], the authors give a synopsis of robust ICS characteristics as well as an ICS cyber resilience evaluation tool. The methods for creating cyber resilience measures are shown in the end, along with an example of how to calculate pliability metrics using the Analytical Hierarchy Process (AHP). A multi-criteria decision support system may be built upon the resilience framework, which can also help technical experts identify areas of unmet research [30–34] need in the field of ICS resilience. In [15], the authors present a strategy for evaluating effectiveness metrics using two decision theory approaches. Analytical Hierarchy Process (AHP) and Analytical Network Process were the MCDM techniques chosen (ANP). Both models make use of numerical scales within an eigenvector-based prioritization technique [35–37].

A security risk assessment is carried out using both qualitative and quantitative methods. All the steps in the risk analysis process were given numerical values in the quantitative cybersecurity risk analysis that was published. To calculate the risk, the analysis is factored into the calculation. This kind of risk assessment is exceedingly intricate, costly, time-consuming, and uses a qualitative approach. The qualitative technique evaluates countermeasures by going through many scenarios for risk probability and threat levels. The qualitative evaluation technique incorporates the assessor's opinions and expertise. The risk is graded on a hierarchical scale, such as critical, high, medium, and low, rather than having numerical numbers in qualitative risk analysis. The focus in this study is on the application of qualitative and quantitative risk assessment approaches [16]. The analytic hierarchy process is also described in [17] as a solution to the problem of accounting information system risk assessment. AHP can, for example, determine the relevance of each option and rank them in order of priority. This is accomplished by placing an overall item on the top level, which is designated by the criteria in the intermediate level. The author suggests an accounting information systems risk assessment algorithm using an analytical hierarchy method. This paper's index is divided into two tiers. Technical risk, operational risk, managerial risk, and emergency risk are the four first-level risk categories.

In [18], the authors used a vulnerability's description to determine priority instead of assigning software flaws a severity level. Convolution neural networks and word embedding are used in this study (CNN). To enable it to recognize distinguishing words and characteristics for the classification job, the CNN is qualified through adequate samples of vulnerability categories across all categories. The authors of [19] propose a technique that predicts the features of software vulnerabilities and then uses those predictions to determine vulnerability severity scores. A dataset of 99,091 records from a sizable, publicly accessible vulnerability database was used to carry out this assignment. In this approach, text analysis and multiple-target classification methods are combined.

Modern enabling methodologies and various computational resources utilized for partitioning and offloading are presented by the authors [20]. Also, they look into applicable paradigms and mobile cloud computing from the standpoint of partitioning and offloading enabling technologies.

After that, they examine a few well-known offloading frameworks and provide a taxonomy for these partitioning and offloading strategies. In hierarchical wireless sensor networks, the authors [21] suggested a brand-new password-based user authentication system, comparing the security and effectiveness of our proposed technique to those of other password-based methods now in use. The authors [22] create LAM-CIoT, a brand-new, lightweight authentication technique for cloud-based IoT environments. An authorized user can remotely view the data from IoT sensors by utilizing LAM-CIoT. LAM-CIoT uses effective "bitwise XOR operations" and "one-way cryptographic hash algorithms. The authors [23] show that the attacker may compute previously established session keys and that the usage of a constant pseudo-identity allows the scheme's user to be traced. Moreover, it presents a unique authentication mechanism for multi-gateway based WSNs rather than attempting to improve a flawed scheme.

Actually, the goal of the Analytic Hierarchy Process (AHP) is to address unstructured issues in management, social sciences, and economics [24]. By using the Analytic Hierarchy procedure (AHP) technique, the decision maker may organize the problem and deconstruct it into a top-down, hierarchical procedure. After that, he or she uses a [25] scale to do a pairwise matrix comparison of the criteria. Eigen vectors or a condensed form with weighted sum (SAW) are used to determine the priorities after normalization.

In contrast, the Evolutionary Optimization Algorithm for Cloud Based Image Retrieval System (EOA-CIRS) approach is introduced by the authors in [26]. During the exhaustive literature review and study of previous related work, it was observed that the prioritization was performed for various data like software vulnerability, accounting information etc. But the ranking of security risk issues and network specific vulnerabilities for Cloud Computing environment is still left. So, there is a need to develop an integrated and dynamic prioritization model which include Cloud Customer interaction, opinions, and reviews. The authors [27, 30] suggested a brand-new user authentication and key management approach. A user and personal server linked to WBAN via the healthcare server housed in the cloud are able to establish a secret session key for their future communication; thanks to the suggested scheme's support for mutual authentication.

Comparing the suggested technique by the authors [21, 28, 29] to other password-based approaches in use, it delivers more security and efficiency. Also, the technique has the advantage of changing the user's password locally and dynamically without the assistance of the base station or gateway node.

24.3 Proposed Method

The Analytical Hierarchy Process is a Multi-Criteria Decision Making (MCDM) process that aggregates criteria specified by the cloud customer using similar and numerical data through the use of normalizing procedures. To arrive at the final ratings or scores for prioritizing Cloud environment vulnerabilities, AHP uses a pairwise technique to evaluate the options in light of a set of criteria. Prioritizing a collection of vulnerabilities that best meet a specified set of requirements in accordance with the needs of cloud consumers is the goal of AHP approaches. A set of standards or independent qualities known as criteria must be met by a number of options that have been accepted via discussion and communication with cloud consumers.

It is necessary to do a security risk assessment that considers to the implementation of academic applications for eLearning universities that are cloud and CC based. IT professionals create specific business processes using the SPI model to fulfil business demands. The ICS-E-Educational deployed University cloud was tested in real time using the suggested technique on 3 distinct platforms: GCP, AWS, and the deployed University's Cloud developed for this work, to identify its weaknesses [8]. 18325 vulnerabilities were found in 2020, and 8350 were found up to the first week of June in 2021. In addition to those, NIST-NVD has currently disclosed more than 34700 vulnerabilities that are unique to the cloud. Among the top CSPs that have revealed their vulnerabilities are GCP-6034, AWS-112&IBM-5142.

Here, the proposed approach is the Analytical Hierarchical Process Model (AHP). In this approach a model is developed based on the priority matrix, which is generated through the Cloud Customer interaction, opinions and reviews. The pairwise matrix comparison of the criteria is performed iteratively using an AHP algorithm. Then on the basis of results of AHP, a new priority matrix is generated which helps the Cloud Security Administrator to analyze and prepare mitigation plans.

Before the implementation of the proposed approach, the authors considered the system having n no. of vulnerabilities as represented by a vulnerability vector [V_1, V_2, V_3 ……. V_n]. These vulnerabilities were taken as input called Alternatives for the proposed algorithm. There are eight Primary Security Criteria based on CVSS are [C_1, C_2 ….C_8] and five Proposed Security Criteria based on threat intelligence as [C_9, C_{10} …. C_{13}]. The details of these criteria are represented in following Table 24.1:

Table 24.1 Primary & proposed security criteria.

Primary security criteria (PSC)	Proposed security criteria (PRSC)
AV: Attack vector	ACC: Number of assets compromised/ critical at all layers.
CA: Complexity of Attack	ATI: Attack Incidents which includes Past threat Patterns, Past threat sources, etc.
RP: Required Privileges	PFD: Patch fixation duration.
UI: User Interaction	DoD: How old/duration of discovery.
SC: Scope	FAI: Frequency of attack incidents.
C: Confidentiality Impact	
I: Integrity impact	
A: Availability Impact	

The vulnerability vector and Primary & Proposed Security Criteria are taken as inputs to the proposed approach based on AHP algorithm. The steps of proposed approach & implementation are as follows:

Step 1: Developing a hierarchical structure with an objective/goal at the top level, the attributes or criteria at the second level and the alternatives at the third Level.

Here, the objective is prioritization of vulnerability which is at the top level of proposed hierarchical structure. The Primary security criteria and Proposed Security Criteria are taken together to develop a single criteria vector as [C1....C13], and placed at second level. The vulnerabilities are represented as alternatives at the third level. The proposed hierarchical structure is represented in the Figure 24.1 as follows:

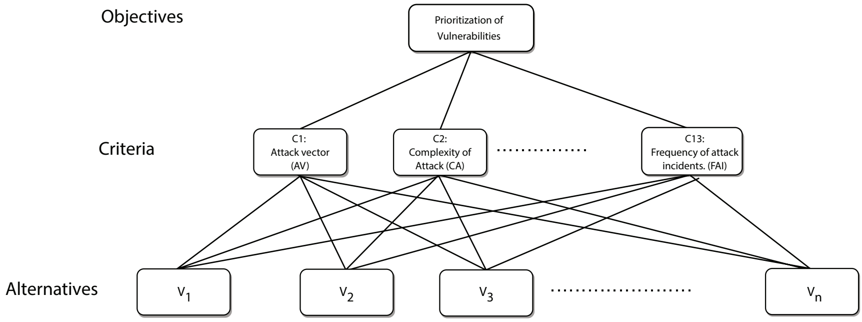


Figure 24.1 Hierarchical structure of proposed approach based on AHP.

Each vulnerability has its own value of all the 13 criteria, and the objective is to prioritize the vulnerabilities represented in the form of vulnerability matrix based on the criteria matrix represented as C1, C2....C13.

Step 2: Determine the relative importance of different attributes or criteria with respect to the goal or objective through creation of a pairwise comparison matrix.

The pairwise comparison matrix represents the relative importance of various attributes or criteria with respect to the objective. One by one criteria pairs are taken and compared based on the relative scale represented in Table 24.2 as follows:

Table 24.2 Scale of relative importance.

1	Equal importance
3	moderate importance
5	strong importance
7	very strong importance
9	extreme importance
2,4,6,8	intermediate values

The pairwise comparison matrix is of size 13x13, which include all 13 criteria and their relative importance. For example, C_1 - C_2 is n12, it means n12 is their relative importance and C_2 - C_1 will be its reciprocal, that is, 1/ n12. Similarly, all the values of relative importance can be filled as shown in Table 24.3:

Table 24.3 Pairwise comparison matrix for security criteria.

Criteria	**C_1**	**C_2**	**C_3**	**C_4**	**C_5**	**C_6**	**C_7**	**C_8**	**C_9**	**C_{10}**	**C_{11}**	**C_{12}**	**C_{13}**
C_1	1	n_{12}	n_{13}	n_{14}	n_{15}	n_{16}	n_{17}	n_{18}	n_{19}	n_{110}	n_{111}	n_{112}	n_{113}
C_2	$1/n_{12}$	1	n_{23}	n_{24}	n_{25}	n_{26}	n_{27}	n_{28}	n_{29}	n_{210}	n_{211}	n_{212}	n_{213}
C_3	$1/n_{13}$	$1/n_{23}$	1	n_{34}	n_{35}	n_{36}	n_{37}	n_{38}	n_{39}	n_{310}	n_{311}	n_{312}	n_{313}
C_4	$1/n_{14}$	$1/n_{24}$	$1/n_{34}$	1	n_{45}	n_{46}	n_{47}	n_{48}	n_{49}	n_{410}	n_{411}	n_{412}	n_{413}
C_5	$1/n_{15}$	$1/n_{25}$	$1/n_{35}$	$1/n_{45}$	1	n_{56}	n_{57}	n_{58}	n_{59}	n_{510}	n_{511}	n_{512}	n_{513}
C_6	$1/n_{16}$	$1/n_{26}$	$1/n_{36}$	$1/n_{46}$	$1/n_{56}$	1	n_{67}	n_{68}	n_{69}	n_{610}	n_{611}	n_{612}	n_{613}
C_7	$1/n_{17}$	$1/n_{27}$	$1/n_{37}$	$1/n_{47}$	$1/n_{57}$	$1/n_{67}$	1	n_{78}	n_{79}	n_{710}	n_{711}	n_{712}	n_{713}
C_8	$1/n_{18}$	$1/n_{28}$	$1/n_{38}$	$1/n_{48}$	$1/n_{58}$	$1/n_{68}$	$1/n_{78}$	1	n_{89}	n_{810}	n_{811}	n_{812}	n_{813}
C_9	$1/n_{19}$	$1/n_{29}$	$1/n_{39}$	$1/n_{49}$	$1/n_{59}$	$1/n_{69}$	$1/n_{79}$	$1/n_{89}$	1	n_{910}	n_{911}	n_{912}	n_{913}
C_{10}	$1/n_{110}$	$1/n_{210}$	$1/n_{310}$	$1/n_{410}$	$1/n_{510}$	$1/n_{610}$	$1/n_{710}$	$1/n_{810}$	$1/n_{910}$	1	n_{1011}	n_{1012}	n_{1013}
C_{11}	$1/n_{111}$	$1/n_{211}$	$1/n_{311}$	$1/n_{411}$	$1/n_{511}$	$1/n_{611}$	$1/n_{711}$	$1/n_{811}$	$1/n_{911}$	$1/n_{1011}$	1	n_{1112}	n_{1113}
C_{12}	$1/n_{112}$	$1/n_{212}$	$1/n_{312}$	$1/n_{412}$	$1/n_{512}$	$1/n_{612}$	$1/n_{712}$	$1/n_{812}$	$1/n_{912}$	$1/n_{1012}$	$1/n_{1112}$	1	n_{1213}
C_{13}	$1/n_{113}$	$1/n_{213}$	$1/n_{313}$	$1/n_{413}$	$1/n_{513}$	$1/n_{613}$	$1/n_{713}$	$1/n_{813}$	$1/n_{913}$	$1/n_{1013}$	$1/n_{1113}$	$1/n_{1213}$	1

Step 3: Calculate the Sum of column of attribute in pairwise comparison matrix.

Let C_{ij} representing the value of ith row and jth column then the sum of column attribute can be calculated iteratively for each column by using following formula:

$$\begin{aligned} &\text{for i=1 to 13} \\ &\{\,\text{Sum[i]} = 1 + \Sigma_{j=1}^{13}\ Cji\,\} \end{aligned} \tag{24.1}$$

Step 4: Normalized pairwise matrix is calculated.

To calculate the normalized pairwise matrix, each value of the matrix will be divided by its respective sum of attributes. It can be given as:

$$\begin{aligned} &\text{for i} = 1 \text{ to } 13 \\ &\text{for j} = 1 \text{ to } 13 \\ &\{\,\mathrm{C_{ij}} = \mathrm{C_{ij}} / \text{Sum[i]}\,\} \end{aligned} \tag{24.2}$$

Step 5: Calculate the criteria weight by averaging the row values in pairwise comparison matrix.

Updated value of matrix row elements which are evaluated through equation no. (24.2) are considered for evaluation of average criteria weight. It can be evaluated for each row by:

$$\begin{gathered}\text{for i=1 to 13} \\ \{\, \mathrm{CW}[\mathrm{i}] = \Sigma_{j=1}^{13}\, Cij \,\}\end{gathered} \tag{24.3}$$

Step 6: Calculate the consistency matrix, again calculate the weighted sum value by summation of each criterion now ratio of weighted sum value with criteria weight. Now calculate $\boldsymbol{\lambda}_{\text{max}}$ through average of this ratio.

Now, to verify the relative importance considered in the step 2 for generation of pairwise comparison matrix, it is required to calculate the consistency matrix, weighted sum and $\boldsymbol{\lambda}_{\text{max}}$ using following formula:

a) Consistency matrix = CW[i] * C_{ij} (24.4)
b) weighted sum of each row $= \mathrm{CW}\,[\mathrm{i}] = \Sigma_{j=1}^{13}\, Cij$ (24.5)
c) for each Row Calculate Ratio, R[i] = weighted sum Sum [i]/ criteria weight CW[i]
d) λ_{max} = average of R[i]
e) Consistency Index (CI) = $\boldsymbol{\lambda}_{\text{max}}$ – n/n-1
f) Consistency Ratio = CI/Random Index (RI)

Where, Random index (RI) is the standard index and as per the following Table 24.4, it is 1.56 for 13 criteria:

Table 24.4 Random index (RI).

n	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
RI	0.00	0.00	0.58	0.9	0.12	0.24	1.32	1.41	1.45	1.49	1.51	1.48	**1.56**	1.57	1.58

Now, check Consistency Ratio with the maximum permissible inconsistency (value = 0.1). If it is under or equal to 0.1, it means the priority matrix is reasonably consistent. Hence, the criteria weights which are generated in step 5 are utilized for prioritization of vulnerabilities. The criteria weight CW[i] will be converted into the weighted percentage and then will be multiplied by the respective value of the criteria of vulnerability for ranking. The final outcome of this approach will be a Priority matrix which will further help the system and security administrator for preparation of mitigation plan.

24.4 Result and Discussion

The following security criteria are taken as input for the proposed AHP as given in Table 24.5: Security Criteria for AHP and in Table 24.6: Pairwise comparison matrix for Security criteria:

Table 24.5 Security criteria for AHP.

id	Security criteria
C_1	AV: Attack vector
C_2	CA: Complexity-of-Attack
C_3	RP: Required-Privileges
C_4	UI: User-Interaction
C_5	SC: Scope
C_6	C: Confidentiality-Impact
C_7	I: Integrity-impact
C_8	A: Availability-Impact
C_9	ACC: Number of assets compromised/critical at all layers.
C_{10}	ATI: Attack Incidents which includes Past threat Patterns, Past threat sources, etc.
C_{11}	PFD: Patch fixation duration.
C_{12}	DoD: How old/duration of discovery.
C_{13}	FAI: Frequency of attack incidents.

Now, as per step 2 and Table no. 24.3, the pairwise comparison matrix will be:

After implementation of step 3 & 4, normalized pairwise comparison matrix will be generated as represented in Table 24.7:

Calculating the criteria weights and percentage according to step 5, the weighted priority matrix will be generated as shown in Table 24.8:

Table 24.6 Pairwise comparison matrix for security criteria.

Criteria	AV	CA	RP	UI	SC	C	I	A	ACC	ATI	PFD	DoD	FA
AV	1.00	3.00	7.00	3.00	3.00	0.33	0.20	0.20	0.20	0.20	0.14	0.14	0.3
CA	0.33	1.00	7.00	5.00	3.00	0.33	0.20	0.20	0.33	0.20	0.14	0.14	0.1
RP	0.14	0.14	1.00	3.00	3.00	0.33	0.20	0.33	0.20	0.20	0.14	0.14	0.1
UI	0.33	0.20	0.33	1.00	5.00	0.33	0.33	0.33	0.33	0.33	0.14	0.14	0.1
SC	0.33	0.33	0.33	0.20	1.00	0.33	3.00	3.00	0.20	0.33	0.14	0.14	0.1
C	3.00	3.00	3.00	3.00	3.00	1.00	3.00	3.00	3.00	3.00	3.00	3.00	3.0
I	5.00	5.00	5.00	3.00	0.33	0.33	1.00	3.00	3.00	3.00	3.00	3.00	3.0
A	5.00	5.00	3.00	3.00	0.33	0.33	0.33	1.00	3.00	0.20	0.20	0.20	0.2
ACC	5.00	3.00	5.00	3.00	5.00	0.33	0.33	0.33	1.00	0.11	0.14	0.14	0.2
ATI	5.00	5.00	5.00	3.00	3.00	0.33	0.33	5.00	9.00	1.00	0.33	0.33	0.2
PFD	7.00	7.00	7.00	7.00	7.00	0.33	0.33	5.00	7.00	3.00	1.00	3.00	0.3
DoD	7.00	7.00	7.00	7.00	7.00	0.33	0.33	5.00	7.00	3.00	0.33	1.00	0.3
FAI	3.00	7.00	7.00	7.00	7.00	0.33	0.33	5.00	5.00	5.00	3.00	3.00	1.0

Table 24.7 Normalized pairwise comparison matrix.

Criteria	AV	CA	RP	UI	SC	C	I	A	ACC	ATI	PFD	DoD	FA
AV	0.023	0.047	0.048	0.051	0.050	0.082	0.023	0.008	0.008	0.011	0.022	0.013	0.0
CA	0.012	0.024	0.071	0.051	0.050	0.055	0.029	0.011	0.013	0.011	0.018	0.013	0.0
RP	0.012	0.008	0.024	0.051	0.050	0.082	0.023	0.014	0.008	0.011	0.027	0.019	0.0
UI	0.012	0.012	0.012	0.025	0.075	0.055	0.039	0.014	0.013	0.019	0.022	0.019	0.0
SC	0.012	0.012	0.012	0.008	0.025	0.041	0.231	0.085	0.008	0.019	0.022	0.019	0.0
C	0.047	0.071	0.048	0.076	0.100	0.164	0.231	0.085	0.076	0.170	0.217	0.189	0.3
I	0.116	0.094	0.119	0.076	0.013	0.082	0.116	0.085	0.114	0.170	0.217	0.189	0.2
A	0.116	0.094	0.071	0.076	0.013	0.082	0.058	0.042	0.076	0.014	0.022	0.019	0.0
ACC	0.116	0.071	0.119	0.076	0.125	0.082	0.039	0.021	0.038	0.006	0.018	0.016	0.0
ATI	0.116	0.118	0.119	0.076	0.075	0.055	0.039	0.127	0.190	0.057	0.036	0.031	0.0
PFD	0.116	0.142	0.095	0.127	0.125	0.082	0.058	0.169	0.152	0.170	0.108	0.189	0.0
DoD	0.163	0.165	0.119	0.127	0.125	0.082	0.058	0.169	0.190	0.170	0.054	0.094	0.0
FAI	0.140	0.142	0.143	0.178	0.175	0.055	0.058	0.169	0.114	0.170	0.217	0.189	0.1

Table 24.8 Weighted priority matrix for security criteria.

Criteria	AV	CA	RP	UI	SC	C	I	A	ACC	ATI	PFD	DoD	FA
%	3.14	2.91	2.69	2.56	3.93	14.20	12.61	5.45	5.79	8.19	12.12	11.99	14.

According to step 6, λ_{max} is calculated for each criterion and with the help of which consistency matrix, consistency index and consistency ratio are also calculated. The value of consistency ratio is 0.1 which is found to be a satisfactory value when compared with the maximum permissible inconsistency value (0.1).

Now, using priority Matrix for security criteria represented in Table 24.8 the priority of the vulnerability represented in Table 24.1 is regenerated. The results are represented in Figure 24.2, shows that

The results of proposed method-1 used for prioritization of vulnerabilities shows that the percentage of critical level is 7.6% and high level is 11.08% for cloud network vulnerability, which was 77% and 2% in previous case. Similarly, the percentage of critical level is 9.41% and high level is 15.13% in case of cloud server vulnerabilities, which was 74% and 6% in previous case. So, after analyzing the result of proposed Method-1 it can be concluded that the percentage of critical level vulnerabilities are reduced up to the acceptable level so that the network and security administrator can successfully design a plan and policy for their mitigation. The following table represents the results in Table 24.9: Comparison of results of Proposed and previous method:

Table 24.9 shows that the percentage of cloud network vulnerabilities was 77% previously and it reduced to 7.64%. Similarly, critical cloud server

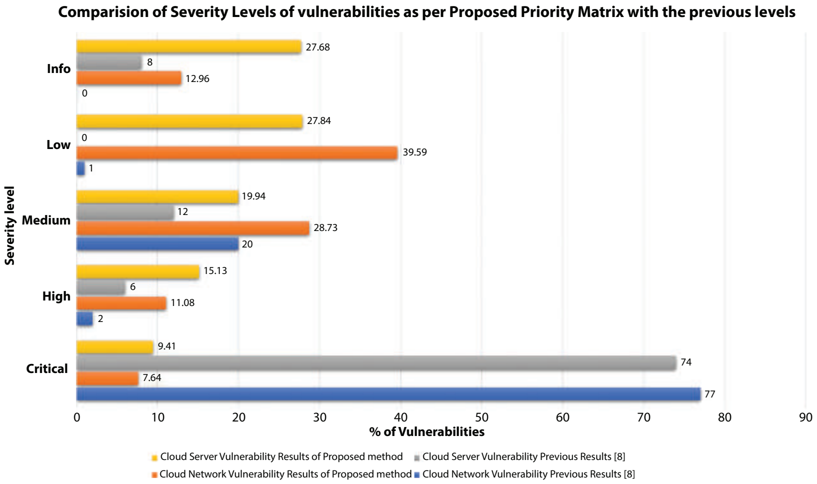


Figure 24.2 Comparison of severity level of vulnerabilities according to the proposed priority matrix with the previous level.

Table 24.9 Comparison of results of proposed and previous method.

Method / Severity		Critical	High	Medium	Low	Info
Cloud Network Vulnerability	Previous Results [8]	77 %	2 %	20 %	1 %	0 %
	Results of Proposed method	7.64 %	11.08 %	28.73 %	39.59 %	12.96
Cloud Server Vulnerability	Previous Results [8]	74 %	6 %	12 %	0 %	8 %
	Results of Proposed method	9.41 %	15.13 %	19.94 %	27.84 %	27.68

vulnerabilities was 74% previously and was reduced to 15.13%. Hence, it helps the Cloud security administrator to decide which vulnerabilities are required to mitigate first and also assist them to develop mitigation plan accordingly.

24.5 Conclusion

It is practically not feasible to resolve every system vulnerability. Therefore, a unique approach and model are proposed for combining AI and machine learning to prioritize security vulnerabilities for Cloud computing environment to achieve the goal. In this work, risk vectors are enhanced by adding more risk variables, and then AI and machine learning techniques are implemented. Five novel characteristics are identified in this research work, and they are combined with the CVSS score to create an enhanced and annotated risk vector. A new risk matrix is then created utilizing this enhanced and updated risk vector system, which will be used as input for both of the suggested options. This work employed the Analytical Hierarchical Process Model (AHP). In this method, a model is created based on the priority matrix that is produced by cloud customer interaction, feedback, and reviews. The AHP technique is used iteratively to compare the criteria in a pairwise matrix. A new priority matrix is then created based on the AHP results, which further will be helpful to the Cloud Security Administrator in the analysis and mitigation plan preparation.

References

1. Sunyaev, A., Cloud Computing, in: *Internet Computing*, Springer, Cham, 2020, https://doi.org/10.1007/978-3-030-34957-8_7.
2. Singh, U.K. and Sharma, A., Cloud computing security framework based on shared responsibility models, in: *Cyber-physical, IoT and autonomous systems in industry 4.0*, 1st ed, V. Bali (Ed.), pp. 39–55, CRC Press Taylor & Francis Group, USA, 2021, https://doi.org/10.1201/9781003146711-3.
3. Alghofaili, Y., Albattah, A., Alrajeh, N., Rassam, M.A., Al-rimy, B.A.S., Secure Cloud Infrastructure: A Survey on Issues, Current Solutions, and Open Challenges. *Appl. Sci.*, 11, 9005, 2021, https://doi.org/10.3390/ app11199005.
4. Sharma, A. and Singh, U.K., Deployment model of e-educational cloud for departmental academic automation using open source. *HTL J.*, 27, 5, 36, 2021, ISSN 1006-6748, https://doi.org/10.37896/HTL27.5/3535.
5. Singh, U.K. *et al.*, Security and Privacy aspect of Cyber-Physical Systems, in: *Cyber physical System: Concept and Application*, 1st ed., CRC press Taylor &

francis group, Chapman and Hall/CRC, USA, 2023, https://doi.org/10.1201/9781003220664-9.

6. Sharma, A. and Singh, U.K., Investigation of Cloud Computing Security Issues & Challenges. *3rd International Conference on Integrated Intelligent Computing Communication & Security (ICIIC), 2021*, 2021, doi: https://doi.org/10.2991/ahis.k.210913.055.
7. Sharma, A., Singh, U.K. *et al.*, An investigation of security risk & taxonomy of Cloud Computing environment. *2021 2nd International Conference on Smart Electronics and Communication (ICOSEC)*, pp. 1056–1063, 2021, doi: 10.1109/ICOSEC51865.2021.9591954.
8. Sharma, A., Singh, U.K. *et al.*, A Comparative analysis of security issues & vulnerabilities of leading Cloud Service Providers and in-house University Cloud platform for hosting E-Educational applications. *IEEE Mysore Sub Section International Conference (MysuruCon)*, 2021, ISBN: 978-0-7381-4662-1.
9. Mishra, A. K., Tyagi, A. K., Dananjayan, S., Rajavat, A., Rawat, H., Rawat, A. (2024). Revolutionizing Government Operations: The Impact of Artificial Intelligence in Public Administration, in: *Conversational Artificial Intelligence*, pp. 607–634, 2024.
10. Silvia Priscila, S. *et al.*, Risk-Based Access Control Mechanism for Internet of Vehicles Using Artificial Intelligence. *Secur. Commun. Netw.*, 2022, 13, 2022, Article ID 3379843, https://doi.org/10.1155/2022/3379843.
11. Current CVSS Score Distribution For All Vulnerabilities, cvedetails.com, https://www.cvedetails.com/ (accessed Sep. 5, 2022).
12. Anjum, M., Kapur, P.K., Agarwal, V., Khatri, S.K., A Framework for Prioritizing Software Vulnerabilities Using Fuzzy Best-Worst Method, pp. 311–316, 2020, 10.1109/ICRITO48877.2020.9197854.
13. Kansal, Y., Kapur, P.K., Kumar, U., Kumar, D., Prioritising vulnerabilities using ANP and evaluating their optimal discovery and patch release time. *Int. J. Math. Oper. Res.*, 14, 236, 2019, 10.1504/IJMOR.2019.097758.
14. Haque, M.A., De Teyou, G.K., Shetty, S., Krishnappa, B., Cyber Resilience Framework for Industrial Control Systems: Concepts, Metrics, and Insights. *2018 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pp. 25–30, 2018, doi: 10.1109/ISI.2018.8587398.
15. Nahar, S., Pithawa, D., Bhardwaj, V., Rawat, R., Rawat, A., Pachlasiya, K., Quantum technology for military applications, in: *Quantum Computing in Cybersecurity*, pp. 313–334, 2023.
16. Petrova, V., The Hierarchical Decision Model of cybersecurity risk assessment. *2021 12th National Conference with International Participation (ELECTRONICA)*, pp. 1–4, 2021, doi: 10.1109/ELECTRONICA52725.2021.9513722.
17. Li-Sheng, Y. and Ru-Ping, D., A Novel Risk Assessment Algorithm for Accounting Information System Using Analytic Hierarchy Process. *2015*

8th International Conference on Intelligent Computation Technology and Automation (ICICTA), pp. 61–64, 2015, doi: 10.1109/ICICTA.2015.24.

18. Sharma, R., Sibal, R., Sabharwal, S., Software vulnerability prioritization using vulnerability description. *Int. J. Syst. Assur. Eng. Manag.*, 12, 58–64, 2021, https://doi.org/10.1007/s13198-020-01021-7.
19. Georgios Spanos, A. and Angelis, L., A multi-target approach to estimate software vulnerability characteristics and severity scores. *J. Syst. Software*, 146, 152–166, 2018, ISSN 0164-1212, https://doi.org/10.1016/j.jss.2018.09.039.
20. Gu, F., Niu, J., Qi, Z., Atiquzzaman, M., Partitioning and offloading in smart mobile devices for mobile cloud computing: State of the art and future directions. *J. Network Comput. Appl.*, 119, 83–96, 2018.
21. Das, A.K., Sharma, P., Chatterjee, S., Sin, J.K., A dynamic password-based user authentication scheme for hierarchical wireless sensor networks. *J. Network Comput. Appl.*, 35, 5, 1646–1656, 2012.
22. Wazid, M., Das, A.K., Vivekananda Bhat, K., Vasilakos, A.V., LAM-CIoT: Lightweight authentication mechanism in cloud-based IoT environment. *J. Network Comput. Appl.*, 150, 102496, 2020, ISSN 1084-8045, https://doi.org/10.1016/j.jnca.2019.102496.
23. Wu, F., Xu, L., Kumari, S., Li, X., Shen, J., Raymond Choo, K.-K., Wazid, M., Das, A.K., An efficient authentication and key agreement scheme for multi-gateway wireless sensor networks in IoT deployment. *J. Network Comput. Appl.*, 89, 72–85, 2017, ISSN 1084-8045, https://doi.org/10.1016/j.jnca.2016.12.008.
24. AVafaei, N., Ribeiro, R.A., Camarinha-Matos, L.M., Normalization Techniques for Multi-Criteria Decision Making: Analytical Hierarchy Process Case Study, in: *Technological Innovation for Cyber-Physical Systems. DoCEIS 2016.* IFIP Advances in Information and Communication Technology, vol. 470, L.M. Camarinha-Matos, A.J. Falcão, N. Vafaei, S. Najdi, (Eds.), Springer, Cham, 2016, https://doi.org/10.1007/978-3-319-31165-4_26.
25. Camarinha-Matos, L.M. and Afsarmanesh, H., Collaborative systems for smart environments: trends and challenges. *Collab. Syst. Smart Networked Environ.*, vol. 434, pp. 3–14, 2014.
26. Vijayakumar, T., Ramalakshmi, K., Priyadharsini, C., Vasanthakumar, S., Sharma, A., Bio-Inspired Optimization Algorithm on Cloud based Image Retrieval System using Deep Features. *2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, Trichy, India, pp. 871–876, 2022, doi: 10.1109/ICAISS55157.2022.10010739.
27. Wazid, M., Das, A.K., Vasilakos, A.V., Authenticated key management protocol for cloud-assisted body area sensor networks. *J. Network Comput. Appl.*, 123, 112–126, 2018, ISSN 1084-8045, https://doi.org/10.1016/j.jnca.2018.09.008.
28. Zekrifa, D.M.S., Sharma, A., Satyam, Patankar, A.J., Bamane, K.D., Data Analyzing with Cloud Computing Including Related Tools and Techniques.

Int. J. Intell. Syst. Appl. Eng., 11, 9s, 233–238, 2023, Retrieved from https://ijisae.org/index.php/IJISAE/article/view/3112.

29. Dixit, V. *et al.*, A Survey on Reliability of Cloud Applications. *Int'l Conf. Emerging Trends and Developments in Science, Management and Technology (ICETDSMT '13)*, pp. 476–481, 2013, ISBN: 978-81-924342-2-3.
30. Zekrifa, D., Sharma, A., Sharma, D., Sharma, R., Rai, S., Pillai, A., A System Based on AI and M1 Enhanced to Investigate Physiological Markers for User Forecasting Decision-Making, pp. 2487–2490, 2023, 10.1109/IC3I59117.2023.10398037.
31. Chirgaiya, S. and Rajavat, A., Tiny object detection model based on competitive multi-layer neural network (TOD-CMLNN). *Intell. Syst. Appl.*, 18, September 2022, 200217, 2023.
32. Rathi, M. and Rajavat, A., *Analysing Cryptographic and Random Data Sanitization Techniques in Privacy Preserving Data Mining*, vol. 83, Allied Publishers, New Delhi, India, 2023.
33. Dhar, S., Dhar, U., Rajavat, A., Factors and Attributes of Team Players: A Study of Engineering Students in India, in: *International Simulation and Gaming Association Conference*, 2021, September, Springer International Publishing, Cham, pp. 53–60.
34. Chirgaiya, S. and Rajavat, A., Comprehensive Analysis of State-of-the-Art Techniques for VQA, in: *Proceedings of International Conference on Advances in Computer Engineering and Communication Systems: ICACECS 2020*, Springer Singapore, pp. 107–115, 2021.
35. Nahar, S., Pithawa, D., Bhardwaj, V., Rawat, R., Rawat, A., Pachlasiya, K., Quantum Technology for Military Applications, in: *Quantum Computing in Cybersecurity*, pp. 313–334, 2023.
36. Pithawa, D., Nahar, S., Bhardwaj, V., Rawat, R., Dronawat, R., Rawat, A., Quantum Computing Technological Design Along with Its Dark Side, in: *Quantum Computing in Cybersecurity*, pp. 295–312, 2023.
37. Chauhan, D., Singh, C., Rawat, R., Chouhan, M., Conversational AI Applications in Ed-Tech Industry: An Analysis of Its Impact and Potential in Education, in: *Conversational Artificial Intelligence*, pp. 411–433, 2024.

25

Cloud Computing Security Through Detection & Mitigation of Zero-Day Attack Using Machine Learning Techniques

Abhishek Sharma[1]* and Umesh Kumar Singh[2]

[1]Shri Vaishnav Vidyapeeth Vishwavidyalaya, Indore, India
[2]Institute of Computer Sciences, Vikram University, Ujjain, India

Abstract

With the extensive use of Cloud computing in areas such as e-government, e-commerce, internet payments, health care, and other daily necessities, the danger of being exposed to numerous threats has grown exponentially. Several businesses are investing in this domain either for their own benefit or as utility to anyone else. The creation of numerous security risks for both producers and society are among the consequences of Cloud-based solutions. The recent global increase in cybercrime and cloud threats has increased the requirements and opportunities for better intrusion detection and prevention systems. The technique used here for enhancing the protection of the cloud computing environment is machine learning (ML). Machine learning techniques have been used in a number of ways on the Cloud to prevent or detect attacks and security issues. But after exhaustive literature review, it was observed that the stand-alone ML model is not sufficient in the current emerging attack techniques. To achieve greater prediction results than would be attained from any of the constituent learning algorithms alone, this work proposes an integrated strategy leveraging ensemble approaches through several learning algorithms and AI technique. Nonmonotonous methodology is also implemented with an ensemble model for enhancing continuous adaptive capabilities of ML model which is further helpful to develop a method for detecting unknown malicious activities on a cloud network for predicting a zero-day attack.

***Keywords*:** Cloud computing, security, machine learning, zero-day attacks, cyber threats, ensemble learning, artificial intelligence

**Corresponding author*: abhiujn9@gmail.com

Rajesh Kumar Chakrawarti, Ranjana Sikarwar, Sanjaya Kumar Sarangi, Samson Arun Raj Albert Raj, Shweta Gupta, Krishnan Sakthidasan Sankaran and Romil Rawat (eds.) Natural Language Processing for Software Engineering, (357–388) © 2025 Scrivener Publishing LLC

25.1 Introduction

The ongoing advancement and sustainable use of the Cloud benefits a multitude of Cloud users in different ways. But as more people utilize the Internet, cloud security becomes more and more important. Cloud security aims to stop unauthorized access and alteration by being closely connected to apps, networking, infrastructure, and diversified data. But as the number of Cloud-connected systems in the financial, e-commerce, e-government, and military domains increases, so does their vulnerability to network attacks, which presents a serious danger and potential for harm. Basically, effective methods for spotting, thwarting, and maintaining security against Cloud attacks are needed. Additionally, different attacks typically require different approaches. How to identify different kinds of cloud attacks raises a fundamental issue with cloud security that needs to be resolved, especially with regard to zero-day attacks, or assaults that have never been seen before. Researchers utilized various types of machine learning approaches to categorize network assaults without existing understanding of their precise features for numerous years. On the other hand, traditional ML methods, are incapable to provide unique features extracted to address the challenge of cloud attack discovery due to model complication/restrictions. Cloud computing presents various issues, including cloud user administration, multi-tenancy support, and application security [1]. The security framework will notify together cloud users and CSPs about clear boundaries and shared duties at every level because security is a key priority in the cloud-based computing systems. By modelling the suggested SRM-based Cloud Computing security framework in their own or other cloud systems, cloud actors may assess security features and compliance [2]. Because of fast technological advancements, many phishing operations have evolved from specific incidents of theft or damage to well-organized and economically sponsored individuals seeking massive gains. In this arena, organized crime's goals vary from economic benefit to attaining diplomatic interests. It shifts forces businesses to explore advanced and complicated methods to remain with evolution of attacks. As an outcome, a newly in demand wave of security solutions, known as Cyber Threat Intelligence is evolving and gaining increasing attention from researchers and protection professionals. It is a cyber-threat information system that delivers substantial proof-based information. Companies can make cyber security decisions based on the information they've gathered, such as identifying, preventing, and healing from breaches [3].

Currently, during COVID-19 pandemic, the frequency of data security assaults has risen considerably since the epidemic has caused enterprises throughout the world to implement work-at-home policies despite implementing necessary and effective countermeasures [4, 12]. Another element of COVID-19 is that hacker groups are seeing an upsurge in discussions about leveraging the epidemic as a new chance for assaults, most prominently assaults hitting distant work tools and outright fraud targeting individuals seeking for employment or knowledge concerning COVID-19 [5]. On the Deep and Dark Webs, social networking sites are seen as vital resources for attackers seeking technical information and honing their abilities. Individuals exchange information, interact with one another, and trade hacking-related products including compromised data and stolen identities over these networks like card numbers, and system vulnerabilities [6]. The number of established Cloud Service Providers (CSPs) is continuously increasing, raising a slew of protection and safety issues. This study [7] examines the security threats and attacks of three tiers of Cyber Physical Systems, explains the security aims and challenges of Cyber Physical Systems, and explains the security objectives and challenges of Cyber Physical Systems. Cloud technology is vulnerable to a variety of security vulnerabilities, ranging from network-level to application-level. Cloud service security identifies the numerous forms of cloud attacks that have happened recently, as well as the effectively adopted ways to limit the risks [8–10].

Cloud computing has evolved into a foundation of the IT industry in recent years, enhancing the capability of virtualization, storing, hosting, and other internet services. However, cyber assaults can compromise the cloud infrastructure. Forty percent of businesses disclosed cloud security vulnerabilities in 2021. Accenture was hit by the LockBit ransomware outbreak in August of 2021. The perpetrators reported to have stolen 6Terabytes of data and demanded a $50 million ransom in exchange. Kaseya, a supplier of IT solutions, discovered an assault on their integrated remote monitoring and trusted network protection application in July 2021. The intruders wanted to take control of Kaseya services from managed service providers to downstream clients. Cognyte, a cyber-analytics business, left a database unencrypted and without authentication mechanisms in May of 2021. As a result, hackers were able to get access to 5 billion records. Names, email addresses, passwords, and vulnerability data points from their system were all compromised. Search engines have even indexed the data. In April 2021, Facebook announced a data breach impacting there were exposed tens of billions of user records publicly on AWS. Two third-party Facebook app development businesses sparked the issue by posting the records in plain sight. The database contained personal data that social engineers may use in targeted attacks or hacking attempts. Online chat software Raychat managed to evade a significant cyberattack in February 2021. After that a cloud database setting was compromised, 267 million usernames, emails, passwords, files, and encrypted chats were exposed to intruders. A targeted bot assault deleted the whole data of the company shortly after that. According to sources, the data was revealed due to a Mongo-DB misconfiguration. This assault proved how bot threat actors may possibly use NoSQL databases as straightforward aims [11].

During April–May 2022, US Federal agencies have been instructed to either immediately patch or temporarily deactivate a set of enterprise products from VMware in response to "active and expected exploitation of multiple vulnerabilities". At the same time, two flaws in the IT monitoring system Icinga's web control interface allowed even unauthorized attackers to execute arbitrary PHP code and take over servers. According to recent research from the DevOps Institute, enterprise IT people believe that cyber security skills are the most crucial technical talents for their teams. Security proficiencies are either "critical" or "important" to the performance of their team's tasks, according to 92% of respondents to the "Upskilling IT 2022" poll. In Feb 2022, a significant cyberattack that interrupted satellite internet in Ukraine and "helped enable President Vladimir Putin's invasion of the nation" was attributed to Russia by the EU. One hour before Russia's invasion of Ukraine on February 24, 2022, an attack on the KA-SAT network knocked down thousands of devices.

The major research objective here is to create a system/model for detecting one or more unknown harmful behaviors on a cloud network to forecast a zero-day assault, which includes a route, a physical layer, or a storage system, among other components.

25.2 Related Work

According to Symantec's 2016 Internet Threat Report [13], targeted assaults have increased by 125 percent over the previous year. In addition, on average, a new zero-day vulnerability was discovered every week in 2015. With 8 zero-day vulnerabilities revealed in 2011, 14 zero-day vulnerabilities reported in 2012, and 23 zero-day vulnerabilities reported in 2013, the number of zero-day vulnerabilities has doubled over the previous year. In 2014, the figure remained stable at 24. However, an increase in zero-day vulnerabilities in 2015 underscores the importance of zero-day assaults.

Up till October 2016, 82 zero-day vulnerabilities had been revealed. Attacks against undiscovered system defects with no patch or repair are known as zero-day attacks [14]. Standard defenses have a hard time detecting zero-day attacks because the conventional security measures depend on malware signs, which are unidentified for zero-day assaults. Invaders are very adept, and its malware may remain undiscovered on systems for many days or may years, giving them enough opportunity to wreak irreversible damage. Trading with unidentified vulnerabilities is obviously a difficult challenge. Even though there are numerous efficient ways for combating known attacks, including as IDS/IPS, firewalls, antivirus, software upgrades, and patching, zero-day assaults are notoriously difficult to defend owing to lack of knowledge. Uncovering unforeseen flaws and determining how to exploit them is clearly a difficult undertaking. The most dangerous of all the threats to an organization's IT infrastructure are zero-day vulnerabilities. They made the faults in the system known to the intruders before a fix was offered. Though zero-day vulnerabilities are unidentified, Software (s/w) vendors may be aware of a defect but have yet to publish a cure. According to a FireEye analysis [15], thieves uncover vulnerabilities that are undisclosed to the community, together with makers, for an average of more than 300 days.

Distributed attack detection is contrasted with centralized detection, and the deep learning model [16] is contrasted with conventional machine learning methods. Our distributed attack detection method beats centralized deep learning-based detection techniques, according to tests. The DL model has also been shown to be more successful at detecting attacks than its shallow counterparts. According to [17], the automobile sector is in the forefront of implementing machine learning algorithms for risk assessment. The most widely used machine learning approach for engineering risk assessment is artificial neural networks. This page also includes additional conclusions from the review process. The authors of [18] create a Honey Net based on Docker technology that collects data to detect and analyze adversary attack patterns. The data is subsequently transformed into pictures, which are then utilized as examples to train a deep learning model. Finally, the AIoT deploys the trained model to identify threats and give situational awareness. The authors of [19] incorporated a dynamic safety risk monitoring, control, and management platform into a building information modeling (BIM) management platform. Finally, the processing and analysis of massive data collected from 3S methods and sensors offer excellent prospects for developing an integrated excavation technology system. According to the professional procedures and standards described in [20], the author used several analytical techniques to generate information based on those standards. The dataset includes about 1940 occurrences of each of the 18 input characteristics. Additionally, it looked at how different machine-learning algorithms behaved to show potential dangers associated with cloud computing. This study describes the impact of testing and training subsets of data by randomly dividing the sub dataset into four different groups. Because the service provider is in charge of ensuring that numerous services are accessible and available, maintenance is inexpensive, and consumers are exempt from issues with resources planning and upkeep on the supplier's side. Cloud Computing has been dubbed "IT on Demand" or "Utility Computing" because of these traits. Scalability is a key element of Cloud Computing, which is achieved by the virtualization of servers [21, 37].

A software fault discovered by attackers before it is discovered by the provider is known as a zero-day vulnerability (ZDVs). Since there is no known fix for ZDVs and no protections are in place on user computers, attacks are extremely likely to be successful. A zero-day incident is a tactic methodology used by threat actors to target organizations with undiscovered vulnerabilities. Zero-day malware, for example, is a malicious software designed by attackers to exploit a ZDVs. ZDAs are harmful acts that entail exploiting an undiscovered product or software vulnerability. Hackers take advantage of them until the manufacturer releases a fix over all affected systems. The programmers distribute applications with known flaws and vulnerabilities initially. The black-hat community finds technology flaws or weaknesses and afterwards runs a 0-day exploit against them. When the developers become aware of such exploits, they create a patch to address the flaws. They distributed a fix to all users to prevent additional attacks because of the issue [22].

The authors of [23] present a robust intelligent technique for detecting ZA signatures. The suggested work is partitioned into two modules: (a) HVA, which uses heavy hitter to produce high volume ZAs, and (b) LVA, which uses graph approach to develop signatures for low volume ZAs. The information is gathered by creating a virtual scenario with ten actual, high-volume attack nodes and three low-volume attack nodes. The binary classification accuracy for real-time assault data is 91.33 percent for binary classification and 90.35 percent for multi-class classification. On the other hand, binary and multi-class classification accuracy for the CICIDS18 benchmark data set is 91.62 percent and 88.98 percent, respectively. In [24], the authors show that in comparison to other attacks, malware attacks have the greatest percentage rate. Malware has become more prevalent, posing a long-term and significant danger to the security of computer systems and the internet. The well-known WannaCry ransomware assault, for example, damaged millions of devices and generated billions of dollars in losses. Every year, the quantity of malware specimens has grown dramatically, with a new malware specimen emerging every 4.6 seconds in 2018. Other forms of attacks and threats discussed in [25, 26] include SQL injection, drive-by attacks, password assaults, man-in-the-middle attacks, authentication attacks, wrapping attacks, watering hole attacks, and web shell attacks. In [27], the author integrates DNN with PCA and introduces IDS to enhance security level and detection latency.

The purpose of the systematic literature review (SLR) presented by the authors in [28] is to give future scholars access to a research resource on the different approaches, strategies, ML, and DL algorithms that have been employed in the detection of zero-day attacks. Since security is a crucial component of systems, the authors in [29] provide systematic instructions as to how businesses can adopt and audit their security. They also give an agile, up-to-date, and state-of-the-art evaluation of intrusion detection systems and their guiding principles. Concrete items used in operational systems are taken into consideration, along with functionalities and development methodologies for these defense mechanisms.

A framework developed by the authors of the current study [30] has the ability to reduce the possibility of zero-click and zero-day attacks on digital transactions. This review is unusual because of its thorough examination of academic sources and creation of a theoretical framework to guard against Zero-Click and Zero-Day attacks on online banks. By lowering the likelihood of these kinds of assaults, the possible application of this framework might greatly increase the security of digital transactions.

A survey of current Intrusion Detection Systems (IDS) for identifying zero-day vulnerabilities is presented by the authors of [31] and is based on the following factors: the sorts of cyberattacks that are employed, the datasets that are used, and the types of network detection systems. Based on the argument that cyberthreats are a public health concern rather than just a competitive factor, the authors of [32] analyze the relationship between the incidence of malware infections in a population of systems and system vulnerabilities. This paper derives structural recommendations for resilience response and policy requirements.

The techniques for identifying zero-day attacks are described by the authors in [33] as Random Forest (RF) and Extreme Gradient Boosting (XGB). The compute architecture that was applied produced better results than earlier techniques that were reported in the IDS literature.

This highlights the need for protecting critical infrastructures from security threats and intrusions. The authors of [33] examined DDoS attacks in SDN-IoT Networks. This article examined several Software-Defined Networking (SDN)-based DDoS attack mitigation techniques to identify the best one for a given set of network needs.

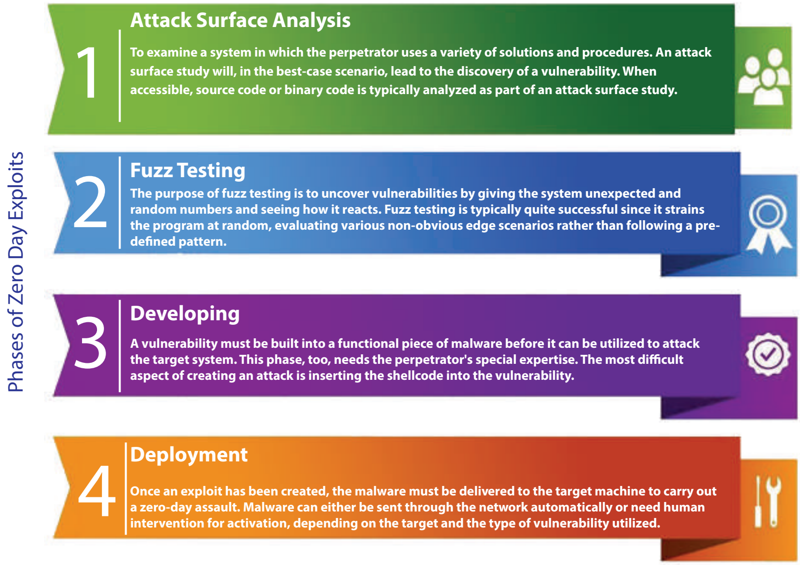
In [34], the author provides a thorough taxonomy and looks at potential ways to detect and prevent intrusions in cloud computing systems. This allows the author to scan, explore, and educate academicians about the most recent IDPSs and alert management techniques. A set of pertinent needs is created, taking into account the intended features of cloud computing and IDPS systems. To meet these requirements, four concepts—fuzzy theory, risk management, ontology, and autonomic computing self-management—are used.

The authors of [35] show notable gains in malware detection rates, with the system effectively seeing new threats that conventional protection methods frequently overlook. The suggested system is a workable and expandable approach that makes use of containerized apps and is easily implementable by small and medium-sized businesses that are eyeing to strengthen their cyber-defense capabilities.

Authors in [36]. The use of container-based security mechanisms for combating zero-day attacks in cloud infrastructures is investigated in this article, which focuses on containerization. The study evaluates how containerization strategies reduce attack surfaces and improve application separation from one another. To forecast and identify key risks as well as threats, the authors of [37–41] conducted risk-based experiments, data analysis, and investigations on IOT and cloud platforms for the purpose of creating mitigation strategies by the security administrator. The model's accuracy rate was 98%. The author of [42] proposed a deep learning approach for detecting and preventing zero-day attacks in the cloud. The HMM model was used in this model to identify attacks, while k-medoids clustering was used to refine attack datasets. In the case of the NSL-KDD dataset, the greatest accuracy was 94%, while in the case of the CIDD dataset, it was 95%.

25.2.1 Analysis of Zero-Day Exploits and Traditional Methods

Before performing actual threat detection in real time, it is advisable to analyze the Zero Day exploits (ZDEs) and how it can be done. Figure 25.1 describes the workflow of ZDEs:



Phases of Zero Day Exploits

1 Attack Surface Analysis

To examine a system in which the perpetrator uses a variety of solutions and procedures. An attack surface study will, in the best-case scenario, lead to the discovery of a vulnerability. When accessible, source code or binary code is typically analyzed as part of an attack surface study.

2 Fuzz Testing

The purpose of fuzz testing is to uncover vulnerabilities by giving the system unexpected and random numbers and seeing how it reacts. Fuzz testing is typically quite successful since it strains the program at random, evaluating various non-obvious edge scenarios rather than following a pre-defined pattern.

3 Developing

A vulnerability must be built into a functional piece of malware before it can be utilized to attack the target system. This phase, too, needs the perpetrator's special expertise. The most difficult aspect of creating an attack is inserting the shellcode into the vulnerability.

4 Deployment

Once an exploit has been created, the malware must be delivered to the target machine to carry out a zero-day assault. Malware can either be sent through the network automatically or need human intervention for activation, depending on the target and the type of vulnerability utilized.

Figure 25.1 Phases of zero day exploits.

Four phases of ZDEs are represented in Figure 25.1. In the first phase, analysis of the attack surface is performed to get the vulnerabilities. During the second phase fuzz testing is performed for uncovered vulnerabilities of the previous phase. During the third phase, development is done. The quantity of storage that is accessible on the exposed stack on the victim machine determines how challenging this assignment is. Code that enables the program to go to the whole executable code should be introduced if there isn't enough free RAM. It's also crucial to have the ability to conceal utilized code and protect it from detection. Recent zero-day assaults involve cutting-edge countermeasures that may be divided into two groups:

- ➢ Metamorphic: Falsification is used to make malware code hard for humans to understand, creating software that is absolutely equivalent but structurally different from the initial.
- ➢ Polymorphic: Encryption methods are employed to encode the source data and bundle that one into encryption algorithm in the payload. Here, Polymorphic zero-day malware is incredibly sophisticated and hard to spot.

During the fifth and last phase deployment is performed. Once an exploit has been created, the malware must be delivered to the target machine to carry out a zero-day assault. Malware can be sent through the web repeatedly or need human intervention for activation, dependent on the target and type of vulnerability utilized. In the latter situation, social engineering techniques like phishing and spam emails are frequently used for persuading customers to copy and execute malware. Mystification and additional techniques to subtly modify the attack sign with every successive assault are utilized by a lot of recent malwares to prevent it from being discovered when used several times. Such measures, which are often taken by offenders when designing malwares, along with the truth that the virus is aiming an undiscovered vulnerability and lacks recognized attack sign to start with, create this type of malware extremely difficult to identify.

Statistics-based detection techniques rely on knowledge of previously identified system vulnerabilities. To compile statistical information on prior assaults and provide a baseline for safe system behavior, machine learning is widely employed in statistics-based detection systems. The primary benefit of these systems is that the more precise they get, the more data they gather. During the course of its operation in a system, a statistics-based solution gathers more data regarding fresh zero-day vulnerabilities, expanding its dataset and creating a more thorough profile for a potential new attack [44].

Though, depending on the standards used, a resolution like this might result in a lot of FP and FN. It might be difficult for programmer developers to strike an appropriate equilibrium with the standards because false negatives must be avoided to prevent missing a zero-day assault, but false positives must be limited to avoid disrupting the company's regular operations. Overall, statistics-based approaches' usefulness in detecting zero-day exploits is restricted. It also has poor detection abilities for malware that is substantially scrambled and disguised. Statistics-based strategies, on the other hand, can be useful for a hybrid solution.

When antivirus software detects malware, it generally uses *signature-based detection* algorithms. The approach, as the name indicates, is based on present malware signature records that are utilized as a guidebook for checking for malware on a system. If often revised, signature repositories are unable to spot new zero-day attacks since by definition, zero-day vulnerabilities lack a verified signature. Therefore, its only way to protect from zero-day assaults using signature-based detection is to use machine learning and related models to create real-time signatures that may meet a virus which is currently unidentified and thus identify it. The following methods can be used to create three different sorts of signatures:

- ➢ Content-based
- ➢ Semantic-based
- ➢ Vulnerability-based signature

Behavior-based detection methods check for malware's features based on how it communicates with the intended system. According to this, a behavior-based approach looks at how incoming files interact with current software rather than the code of those files to determine if such interactions are the result of malicious activity. Building baseline behavior based on information from prior and ongoing system interactions is typically done using machine learning. Similar to statistics-based detection methods, this identification becomes more reliable when more data are provided. A behavior-based detection method which runs for long period on a single target system might be highly good in predicting the outcomes of present operations and detecting malicious software [44].

Hybrid detection approaches are designed to use various capabilities of the three techniques outlined above while preventing their flaws. Hybrid detection systems typically combine 2 or 3 approaches to generate more precise findings. A statistics-based method, for example, may be used to reinforce a behavior-based standard for common behavior and accelerate learning method, while signature-based technique can be applied to filter false positives and improve detection accuracy.

Through the presented and exhaustive work done, it is observed that any single approach, technique or model is unable to work properly or may be less effective because it will not include various Cloud exploit scenarios. Hence, the authors strongly propose an integrated and smart model or framework to perform analysis and detection of Zero-day exploits. It is also recommended that a specific mitigation plan or policy be proposed to reduce the effect of attacks.

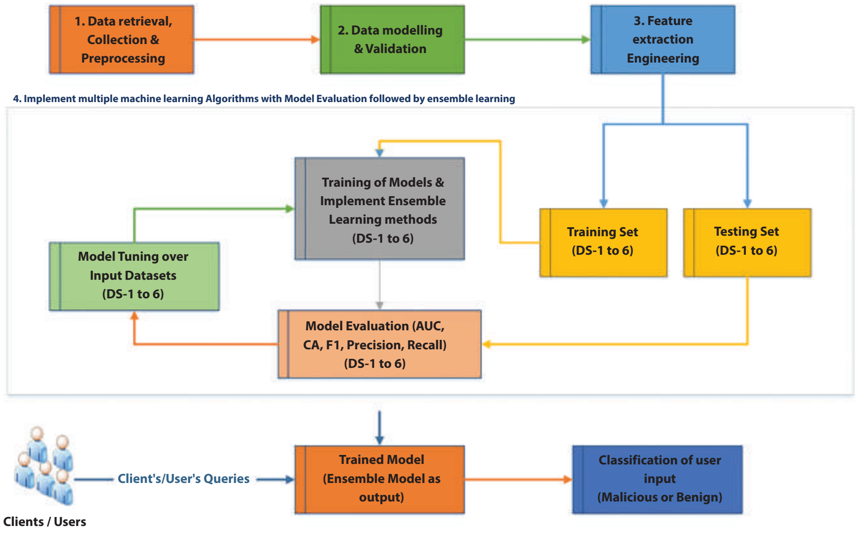
25.3 Proposed Methodology

To develop a method for detecting unknown malicious activities on a cloud network in its ability to forecast a zero-day attack, as well as to analyze and mitigate potential losses in cloud platforms and security of an enterprise-wide cloud in terms of operational, informational, and security networks, proper security risk investigation is required. An integrated strategy is created utilizing different machine learning algorithms simultaneously to investigate security threats in terms of threads and identify intrusion or detection of assaults during the runtime. It is necessary to collect data from memory dumps and log files and prepare various data sets prior to deployment. The data sets are then preprocessed using a variety of ways to avoid under and over fitting. The connection between the various attributes must be discovered during the preprocessing of the Data set [51–53] for it to be free of outliers. Before implementation of the proposed method, it is required to define Security objectives and requirements for zero-day protection and design ensemble learning and predictive models on the basis of them. Figure 25.2 presents the Road map for training algorithms and classification of new customer connection:

The following steps show the proposed approach and methodology:

Step 1: Data retrieval, Collection & Preprocessing

Data retrieval and data collection is the backbone of any machine learning model, that's why in this step, six different datasets are taken as an input based on various threats and risks. After data collection from various sources, the dump files and system files are examined and then csv files are prepared with respective Features [54–56]. During data preprocessing, data cleaning, integration, transformation, reduction are performed. Then preprocessed data sets are digitized, so that they will reduce the size of dataset file and time complexity during the implementation of next level machine learning algorithms. The process will be described as shown below:



1. Data retrieval, Collection & Preprocessing
2. Data modelling & Validation
3. Feature extraction Engineering
4. Implement multiple machine learning Algorithms with Model Evaluation followed by ensemble learning
Training of Models & Implement Ensemble Learning methods (DS-1 to 6)
Training Set (DS-1 to 6)
Testing Set (DS-1 to 6)
Model Tuning over Input Datasets (DS-1 to 6)
Model Evaluation (AUC, CA, F1, Precision, Recall) (DS-1 to 6)
Clients / Users
Client's/User's Queries
Trained Model (Ensemble Model as output)
Classification of user input (Malicious or Benign)

Figure 25.2 System flow for training and classification of cloud network traffic.

a) Collect Data from log files, memory dump files, and miscellaneous files.
b) Present the gathered data into csv format.
c) Digitize all the features of the dataset using numeric values.
d) Define and declare the target feature within the proposed dataset.
e) Fill up the missing parameters of the dataset by finding 'mode' of the respective feature.

Based on the various datasets six different models will be constructed for the ensemble machine learning model. The following data sets are collected as given in Table 25.1: Specification of Data Sets used and their respective threats/risk:

Results: The outcome of this step is Six Data set csv files, which are used as input to the next step.

Step 2: Data modelling and Validation

The following are the steps performed to get the desired dataset csv file before the implementation of feature extraction and selection:

a) Import the processed datasets within the experimental setup built on GCP using Jupyter Notebook.
b) Visualize the finalized digital dataset having no missing values and outliers for each dataset.
c) Define and declare target features in each dataset.

Result: Finalized digital dataset is imported.

Step 3: Feature extraction Engineering

After visualization of the datasets received from the previous step, it is observed that they all are derived from random sampling methods and the parameters are continuous and normally distributed, so during the feature extraction engineering, it is recommended that Pearson Correlation coefficient be used following this sequence:

a) Find the Pearson Correlation coefficient of all the pairs of features for each dataset using following expression:

$$R = \frac{n(\Sigma\alpha\beta) - (\Sigma\alpha)(\Sigma\beta)}{\surd(n\Sigma\alpha^2 - (\Sigma\alpha)^2)(n\Sigma\beta^2 - (\Sigma\beta)^2)} \tag{25.1}$$

Table 25.1 Specification of Data Sets used and their respective threats/risk.

Data set	Distribution of data set	Threat/risk	Reference
Malware Memory Analysis (CIC-MalMem-2022)	Spyware: 10020 Ransomware: 9791 Trojan Horse: 9487 Benign:29298	Malware (Spyware, Ransomware, Trojan Horse)	[43]
Malicious Domains using DNS Traffic Analysis (CIC-Bell-DNS 2021)	Spam: 4337 Phishing: 5001 Malware: 5001 Benign:500001	Spam, phishing, and malware	[44]
Darknet traffic classification (CIC-Darknet2020)	Darknet: 24311 Benign: 134348	TOR, VPN	[45, 46]
DDoS attack (CICDDoS2019)	Portmap:186950 BENIGN: 4734	PortMap; NetBIOS; LDAP; MSSQL; UDP; UDP-Lag, SYN; NTP; DNS; SNMP; SSDP;WebDDoS; TFTP	[47]
IDS (CSE-CIC-IDS2018)	Benign:13484755 DDoS:1263902 DoS:654352 Brute force:380988 Botnet:286188 Infiltration:161843 Web attack:974	Brute-force, Heartbleed; Botnet; DoS; DDoS; Web attacks; infiltration of the network from inside;	[48]
ISOT -2020	Benign: 25076703 Malicious: 11862282	HTTP Flood DOS; Dictionary/Brute Force login attacks; NetworkScanning;DNS Amplification DOS;Synflood DOS; UDP Flood DOS;Probing; Backdoor (reverse shell); Remote-to-Local (R2L);Network Scanning; Trojan Horse; Unclassified (unsolicited traffic);	[49, 50]

b) Identify the features on the basis of correlation threshold value for each dataset.
c) Find the outliers and discard them in each dataset.

Result: Final data set with selected features in csv files.

Step 4: Implementation of multiple machine learning Algorithms with Model Evaluation and tuning over all six datasets followed by ensemble learning

Before the implementation of multiple machine learning algorithms over all six datasets parallelly using experimental setup prepared through Jupyter notebook over GCP, splitting the data set is performed and training and test datasets are identified in the ratio of 75-25 %. Following are the steps:

a) Implementation of K-nearest neighbor (kNN):

The kNN classifier often starts with the Euclidean distance for test and chosen training samples. Let x_a be an input sample with f features (a=1to n), p:no. of features (b=1 to f), and x_{a1}, x_{a2}, ..., x_{af} be the no. of input samples. The following equation describes the Euclidean distance for sample x_a and x_c (c=1to n):

$$\mathrm{d}(\mathrm{x}_a, \mathrm{x}_c) = \sqrt{(x_{a1} + x_{c1})^2 + (x_{a2} + x_{c2})^2 + \cdots .. + (x_{af} + x_{cf})^2} \quad (25.2)$$

If the dependent variable and one or more independent variables are present, the 1-nearest neighbor rule sets the test piece x's projected class to be identical to the true class I of closest neighbor of them, where I is x's closest neighbor:

$$\mathrm{d}(\alpha_i, \mathrm{x}) = \min_b\{\mathrm{d}(\mathrm{m}_b, \mathrm{x})\} \quad (25.3)$$

Model parameters used:
Number of neighbors: 5; Metric: Euclidean; Weight: Uniform;

b) Implementation of Support vector machine (SVM):

It is the most suitable supervised ML technique. Both classification and regression might be accomplished using SVM. The method may be trained using labeled data, and it can produce a hyperplane-based classification of the data into classes that optimizes margin across all attack classes. The SVM, which functions as a binary classifier, can also conduct multi-class categorization in a cascade-style fashion. SVM mostly depends on the settings and types of kernels utilized. The Gaussian kernel calculated using a support vector (SV) is an exponentially declining function in the source feature space having highest value at the SV and uniform decay in all dimensions around the SV, resulting in hyper-spherical edges of the kernel function. The Gaussian or RBF kernel is:

$$K(\alpha,\beta) = EXP(-\frac{||\alpha-\beta||^2}{2\sigma^2}) \tag{25.4}$$

Here, ||α-β|| represents the Euclidean distance. The linear weighted arrangement of the kernel function generated within a data point and every SV is the SVM classifier having Gaussian kernel. The importance of SV in data point categorization is tempered by the SV's global prediction utility, and K (α, β), the SV's local influence on prediction at a specific data point. The Model parameters used are:

- ➢ SVM type: - SVM, C=1.0, ε=0.1;
- ➢ Kernel: - Sigmoid, tanh (auto α·β + 1.0);
- ➢ Tolerance (Numerical): 0.001;
- ➢ limit of Iteration: 100;

c) Implementation of Random Forest (RF):

It is a sophisticated non-linear supervised method employed for regression and classification. This will create several decision trees while training the model, since results of predictions from each tree is combined to produce single conclusion This is referred to as an ensemble approach. The way RF classifiers operate is as observed: the more trees in the model, the greater the accuracy and less likely the model is to be overfit. A collection of decision trees with controlled variance is also made with it. Random forest builds multiple decision trees instead of just one to predict the final output activity class of mobile phone users based on their phone log data. By generating many decision trees for a given dataset, it mitigates the over-fitting problem caused by the one decision tree that was initially supplied. The model parameters that are employed are:

- ➢ Number of trees: 10
- ➢ Maximal number of considered features: unlimited
- ➢ Replicable training: No

- ➢ Maximal tree depth: unlimited
- ➢ Stop splitting nodes with maximum instances: 5

d) Implementation of logistic Regression (LR):
To examine the distinct collection of classes, LR is a supervised machine learning (ML) classification technique. The cost function, often known as the sigmoid function, is used by the logistic function. Predictions are converted to probabilities using this function. It is also suggested that the likelihood of an event occurring may be anticipated by fitting data to the logistic function. Logistic regression computes probabilities by utilizing a logistic function, which is also referred to as a sigmoid function. The function is often restricted to a range of 0 to 1, according to the logistic regression hypothesis. This classifier evaluates the relationship between several independent variables and the category dependent variable for a given dataset. Model parameters used for Regularization: Ridge (L2), C=1, class weights=False.

e) Implementation of Naive Bayes (NB) algorithms:
This model is based on a probability of set of Bayesian equations:

$$P(\alpha|\beta)=P(\beta|\alpha)*P(\alpha)/P(\beta) \tag{25.5}$$

This presupposes feature independence and the capacity to consider multiple attributes. Where $P(\alpha)$ is the chances that event α will occur, $P(\beta)$ is the chances that incidentβ will happen, $P(\alpha|\beta)$ is the chances that α will happen given β, $P(\beta|\alpha)$ is the chances that β will happen given α, and $P(\alpha\beta)$ is the chances that α and β will happen. It is assumed that the Naive Bayes properties are not related to each other. The classifier then determines the likelihood of every class until selecting the one with the greatest probability.

f) Implementation of Neural Network (NN):
NN is a type of distributed computing with biological inspiration. It is made up of connections between basic processing units, or nodes. Any two units that are connected have random weights, and they are used to estimate the degree to which one unit will impact the other. The unit serves as input nodes and output nodes, respectively, performing summing and thresholding. The over-6tting issue is typically solved by an early halting method. The early halting technique uses a different validation set to monitor network performance. As the network learns the data, the errors in the

validation set frequently decrease, then increase as the network discovers the peculiarities in the training data. A forward pass and a backward pass are the two steps of a feed-forward neural network. In the forward pass, the network is given a sample input and activations are allowed to flow until the output layer is reached. The linear sum, sigmoid function, and Gaussian function, ReLu, are 3 activation functions that are often used. During the backward pass, the network's actual output (from the forward pass) is contrasted with the desired output, and inaccuracy assessments are computed for the output nodes. To lessen the inaccuracies, one might change the weights that are attached to the output units. The Model Parameters used are:

- Hidden layers: 100;
- Activation: ReLu;
- Alpha: 0.0001;
- Max iterations: 200;
- Replicable training: True;
- ReLU (Rectified Linear Units) function, h(α) = Max (α, 0);

g) Evaluation of all the supervised classification algorithms or models with respective performance matrix:

Five performance metrics are developed in this research work using the confusion matrix: AUC, Accuracy, F1, Precision, and Recall. AUC stands for Area under the receiver operating characteristic curve (ROC). The accuracy metric is determined by dividing all true forecasts by all projected values, which includes all accurate predictions.

$$Accuracy\ (CA) = \frac{(TP_i + TN_i)}{(TP_i + TN_i + FP_i + FN_i)} \tag{25.6}$$

where the instances of the result data designated as TP, TN, FP, and FN are, respectively, true positive, true negative, false positive, and false negative. By dividing the true positive values between the genuine positive values and the false positive values, the accuracy measure is derived.

$$\text{Precision} = \frac{(TP_i)}{(TP_i + FP_i)} \tag{25.7}$$

where TP and FP stand for the result data's true positive and false positive cases, respectively. The recall measure is computed in a manner similar to how the accuracy measure is obtained.

$$Recall = \frac{(TP_i)}{(TP_i + FN_i)} \tag{25.8}$$

where TP and FN stand for the result data's true positive and false negative situations, respectively. F1 will be evaluated as:

$$\text{F1Score} = \frac{2 * (Precision_i * Recall_i)}{(Precision_i + Recall_i)} \tag{25.9}$$

The classification outcomes of the variations utilized in this article will be evaluated using these five-performance metrics. A new measure that will be covered in the following part will be made using the entire set of these five-performance metrics.

Results: Performance Matrix content AUC, accuracy, F1, Precision and recall.

This machine learning model (MLM) approach mixes many models with various datasets to get the best prediction model. The primary concept underlying the ensemble technique is to combine all weak learners into one strong learner, increasing the model's accuracy. Bagging, Boosting, and Stacking are a few methods of frequent ensemble technique types. The bagging ensemble approach uses the partial decision tree algorithm (PART), Adaptive Boost, and Naive Bayes to figure out the trail. It demonstrated that the ensemble technique outperforms PART, Naive Bayes, and Adaptive Boost.

Step 5: Update predictive Set of belief using non monotonic reasoning (AI technique)

According to the performance matrix received from the previous step as output, the model was built for deployment for real time prediction. When the proposed model will be deployed over the Cloud Computing environment, the cloud network traffic will be monitored, then this data will be considered as input for prediction.

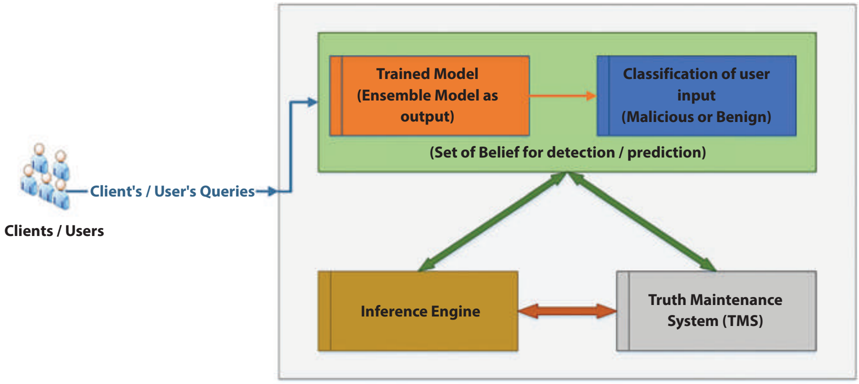

Trained Model
(Ensemble Model as output)
Classification of user input
(Malicious or Benign)
(Set of Belief for detection / prediction)
Client's / User's Queries
Clients / Users
Inference Engine
Truth Maintenance System (TMS)

Figure 25.3 Adaptive Predictive Ensemble Machine Learning (APEML) system.

The outcome of prediction is then taken as input for maintaining a set of beliefs or knowledge base, which will be further used for enabling adaptive and predictive ensemble machine learning models (APEML model) to implement AI techniques like non-monotonous reasoning. Figure 25.3 will present the process as given below.

In Figure 25.3, the set of beliefs represents knowledge base as input to the proposed model. The set of beliefs will adaptively update itself with the arrival of new Cloud Computing environment network information through large numbers of new user's queries. Then the results of newly prediction values will also be included in the adaptive set of beliefs in the form of a knowledge base using non monotonous reasoning. The Truth Maintenance system is supposed to update the set of beliefs on the basis of previous prediction and the new client's queries.

25.4 Results and Discussion

Table 25.2 encapsulates the performance parameters of all six datasets in terms of AUC, Accuracy, F1, Precision and recall achieved by each algorithm. It is evident that from the confusion matrix shown in Table 25.2, the maximum accuracy (CA) achieved on DS-1 is 1 in case of RF, NN and second highest is .99 in case of kNN, NB and LR. The average area under ROC curve (AUC) and average accuracy (CA) is 0.986. Whereas average F1, precision and recall value is 0.983. The confusion matrix of all six datasets is given in Table 25.2:

Table 25.2 Confusion matrix of all six datasets with six algorithms/classifiers.

Dataset/ model	Method/ algorithm	Evaluation of results: confusion matrix				
		AUC	**CA**	**F1**	**Precision**	**Recall**
DS-1	kNN	0.99	0.99	0.99	0.99	0.98
	SVM	0.98	0.97	0.79	0.82	0.76
	RF	1	1	1	1	1
	NN	1	1	0.99	0.99	0.99
	NB	0.99	0.99	0.97	0.93	1
	LR	0.99	0.99	0.991	0.996	0.97
DS-2	kNN	0.995	0.997	0.998	0.997	1
	SVM	0.936	0.891	0.941	0.914	0.969
	RF	1	1	1	1	1
	NN	1	1	1	1	1
	NB	0.999	0.986	0.992	0.998	0.986
	LR	1	1	1	1	1
DS-3	kNN	0.92	0.97	0.98	0.98	0.99
	SVM	0.97	0.97	0.98	0.98	0.98
	RF	1	1	1	1	1
	NN	1	1	1	1	1
	NB	0.99	0.98	0.99	0.99	0.99
	LR	0.98	0.97	0.98	0.99	0.98
DS-4	kNN	0.99	0.99	0.99	0.99	0.99
	SVM	0.97	0.95	0.95	0.95	0.95
	RF	1	1	1	1	1
	NN	1	1	1	1	1
	NB	0.99	0.96	0.96	0.97	0.96
	LR	0.99	0.99	0.99	0.99	0.99
DS-5	kNN	1	1	1	1	1
	SVM	0.96	0.94	0.94	0.94	0.94
	RF	1	1	1	1	1
	NN	1	1	1	1	1
	NB	0.99	0.99	0.99	0.99	0.99
	LR	0.99	0.99	0.99	0.99	0.99

(*Continued*)

Table 25.2 Confusion matrix of all six datasets with six algorithms/classifiers. (*Continued*)

Dataset/ model	Method/ algorithm	Evaluation of results: confusion matrix				
		AUC	**CA**	**F1**	**Precision**	**Recall**
DS-6	kNN	1	1	1	1	1
	SVM	0.887	0.998	0.998	0.998	0.998
	RF	1	1	1	1	1
	NN	1	1	1	1	1
	NB	1	0.99	0.99	0.99	0.99
	LR	1	1	1	1	1

In Figure 25.4, A comparative analysis of performance of machine learning algorithms for each input data set in terms of accuracy is presented. The graph presents the variation of accuracy level of the kNN algorithm from 99 - 100% with respect to the data set used. Similarly, the accuracy level of SPM is observed from 89.1% to 99.8%. Accuracy of Random Forest and neural network is observed as hundred percent in all cases is given in Figure 25.4.

In Naive Bayes (NB) algorithm, the accuracy level is from 96% to 99%. But in case of Logistic regression the accuracy level is almost hundred percent in every case indicated in Figure 25.4.

Figure 25.5 depicts the trade-off between sensitivity (or TP) and specificity using the ROC curve (1 – FP). Classifiers that give curves that are

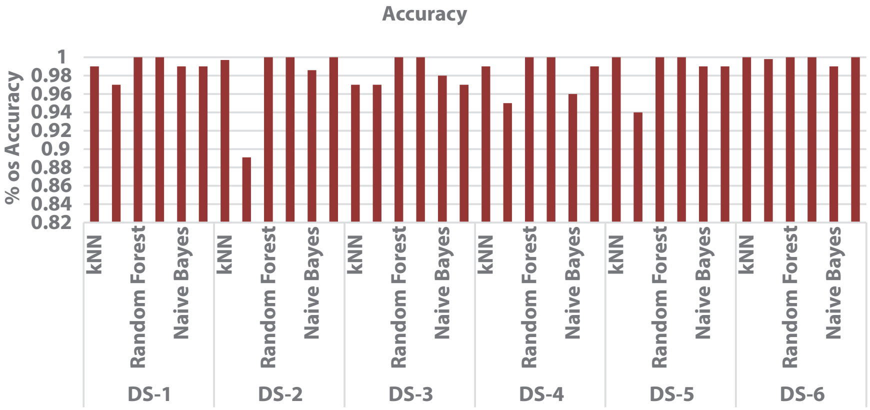
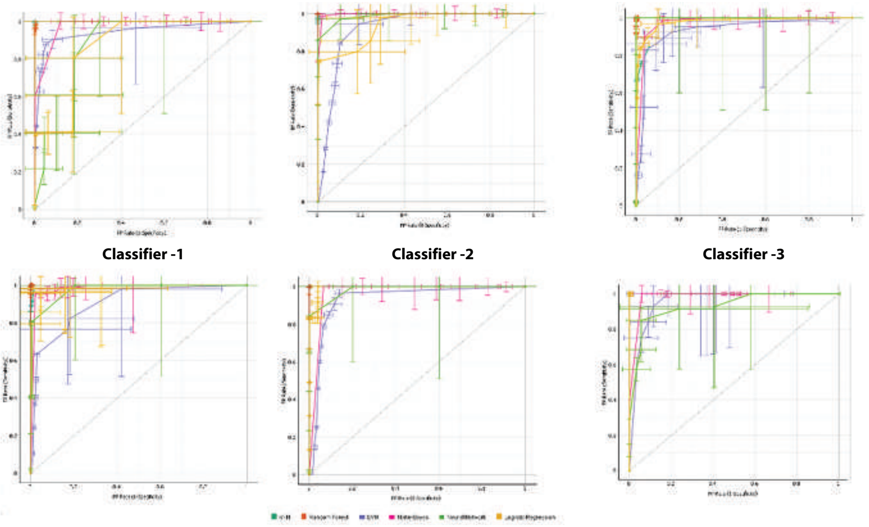


Figure 25.4 Comparative analysis of accuracy for ML algorithms for input datasets.



Classifier -1

Classifier -2

Classifier -3

Classifier -4

Classifier -5

Classifier -6

Figure 25.5 Area under ROC curve for classifier 1 to 6.

nearer to the top-left corner exhibit superior performance. It expects a random classifier to provide diagonal points (FP = TP) as a starting point. An AUC of 0.5 often indicates no classification, whereas values between 0.7 and 0.8 are regarded as acceptable, between 0.8 and 0.9 as outstanding, and values beyond 0.9 as remarkable. The target probability for target class 1 is 89% in case of DS-1, 90% for DS-2, 93% for DS-3, 94% for DS-4, 91% for DS-5 and 91% for DS-6 as shown in Figure 25.6. High-dimensional data are analyzed using the principal component analysis (PCA) in the framework of DS-1 to DS-6. Finding a smaller set of features that accurately represents the original data in a lower-dimensional subspace with little information loss is the purpose of PCA usage here.

In Figure 25.6, the result of PCA for classifier 1 to 6 is presented. The results show the portion of variance with the principal component which includes cumulative variance and component variance for each data set. The explained variance for DS-1 to DS6 is 71%, 53%, 60%, 44%, 45%, and 83% respectively.

The proposed methodology is able to identify 12 different kinds of attacks including frequently used attacks like Malware, ransomware, DDOS etc. The percentage of attack detection using proposed methodology is shown in Table 25.3.

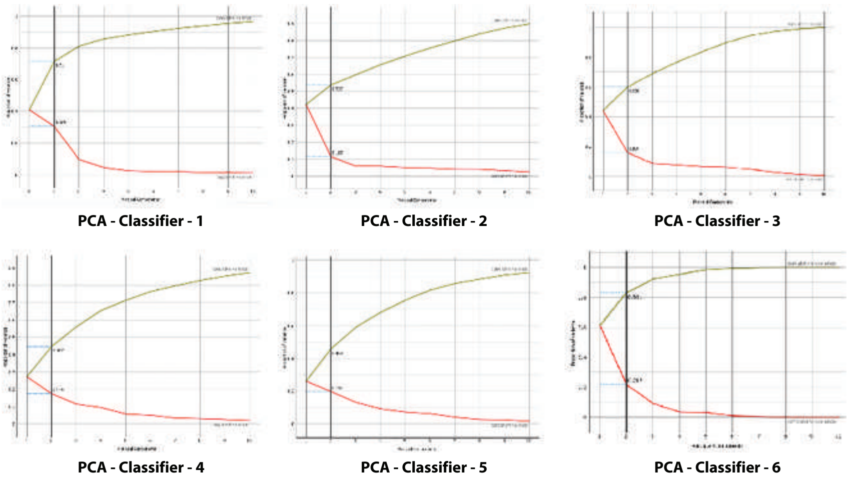

PCA - Classifier - 1
PCA - Classifier - 2
PCA - Classifier - 3
PCA - Classifier - 4
PCA - Classifier - 5
PCA - Classifier - 6

Figure 25.6 Principal component analysis for classifier 1 to 6.

Table 25.3 Detection of top 12 attacks/risk using proposed methodology

S. no.	Attack type	Accuracy/ detection (%)
1	Botnet attack	94%
2	Brute force Attack	94%
3	DoS/DDoS + PortScan	99%
4	Infiltration attack	97%
5	Malware	97%
6	Phishing	98%
7	Ransomware	97%
8	Spam	97%
9	Spyware	98%
10	Trojan Horse	97%
11	VPN and Tor (Darknet)	98%
12	Web attack	96%

Table 25.4 Comparison of performance of proposed model & previous methods/models.

S. no.	Method/model	Accuracy (%)	References
1.	Signature Extraction method	91.33	[23]
2.	HMM	94.00	[42]
3.	CNN for Darknet traffic	86.00	[46]
4.	Proposed APEML Model	99.99	Proposed

During the comparison of performance in terms of Accuracy (CA) of proposed APEML Model and existing methods, it is found that the average CA of the proposed Model is the best among the previous methods. The details are as shown below in Table 25.4.

25.4.1 Prevention & Mitigation of Zero Day Attacks (ZDAs)

Zero-day vulnerabilities and their accompanying assaults will grow indefinitely. If human error and supply chain attack surfaces continue, businesses should set up and meticulously monitor the tools, procedures, and policies to reduce the risk. Figure 25.7 depicts ten such techniques that can aid in the management of zero-day risk.

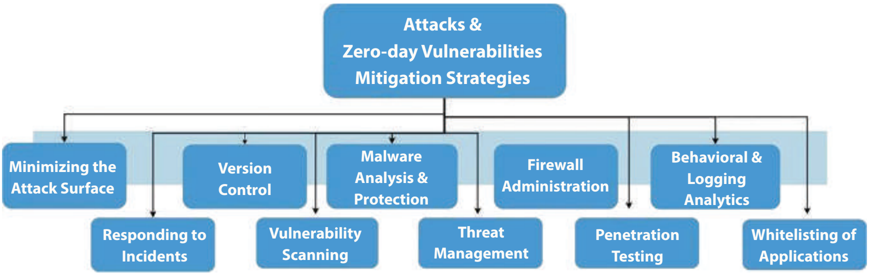

Attacks & Zero-day Vulnerabilities Mitigation Strategies
Minimizing the Attack Surface
Version Control
Malware Analysis & Protection
Firewall Administration
Behavioral & Logging Analytics
Responding to Incidents
Vulnerability Scanning
Threat Management
Penetration Testing
Whitelisting of Applications

Figure 25.7 Top 10 zero day mitigation strategies.

- Attack Surface Reduction (ASR), which prevents malware from infecting computers by blocking attacks that come from scripts, Office files, and emails. ASR can provide favorable conditions while stopping the underlying actions of dangerous documents. It can identify and terminate malicious PowerShell scripts, JavaScript, VBScript, and macro code. It can also prevent the execution of executable email content and payloads downloaded from the Internet.
- Exploit Guard's network security prevents malware from connecting to a command-and-control server by blocking all outbound links before they are used. Outbound network traffic is analyzed using hostname and IP reputation, and connections to unreliable locations are severed.
- Controlled folder access keeps track of the modifications that programs the data in secured directories. Important directories may be secured, and only approved applications are given access. It could stop ransomware from encrypting data.

The Best Practices for Defense Against Zero-Day Attacks are as follows:

a) Implement Patch Management
b) Incident Response Plan should be ready.

The strategy should also include the following:

- The following should also be part of the plan: ✚ Prepare by doing a risk analysis and identifying the most critical assets, which the security team should focus on. Write a document describing the roles, duties, and protocols.
- Identification: describe how to spot a possible ZDA, verify that it's an occurrence, and define what else has to be found out to deal with the threat.
- When a security flaw is identified, containment refers to the immediate activities that can be done to prevent further damage and the longer-term steps that can be taken to rebuild and clean affected systems.
- Eradication: identifying the main cause of the attack and taking steps to prevent it from happening again.

- Recovery: how to restart, inspect, and monitor production systems to ensure that everything is running normally.
- Learnings: No later than two weeks following the attack, do a retrospective to review organizational policies and tooling and figure out how to be suitably prepared for the next attack.

Protection from dangers including 0-day assaults, advanced persistent threats (APT), sophisticated malware & trojans that could avoid conventional signature-based security procedures is provided by the advanced threat detection and response platform. Continuously monitoring of the Vulnerability scanning reports, Patch management, Input validation and sanitization should also be performed for prevention and mitigation of Zero-day Attacks.

25.5 Conclusion and Future Work

The proposed and current methodologies and outcomes are compared in this paper along with a thorough analysis. Each dataset's preparation step's data cleaning procedure was described. To lower the size of the dataset for effective analysis, the dimension reduction is also carried out via correlation. Each dataset is then subjected to an application of ensemble ML algorithms, and the outcomes are examined. The efficacy of identifying intrusions using the suggested dataset and ensemble ML algorithm is 99%, which is superior to the previous findings, according to the trials carried out and assessment results. The accuracy is nearly 100% when utilizing the kNN, RF, and NN techniques, which is highly acceptable. The contributions of this work are as follows:

- This article presents a detailed analysis and comparisons of various Datasets with 12 different categories of attacks/risk.
- This article also presents a detailed implementation of proposed method for evaluation of five performance parameter like Accuracy, AUC, F1, precision and recall.
- This article details how to develop a method for detecting unknown malicious activities using ensemble ML algorithm on a cloud network which adaptively evolve from the proposed system using non-monotonous reasoning for predicting a zero-day attack.

- This article also analyzes and mitigates business risks in cloud network and security of an enterprise-wide cloud with respect to Operational and Security networks to maintain a high level of integrity and security.

References

1. Karthiban, K. and Smys, S., Privacy preserving approaches in cloud computing, in: *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, pp. 462–467, 2018.
2. Singh, U.K. and Sharma, A., Cloud computing security framework based on shared responsibility models, in: *Cyber-physical, IoT and autonomous systems in industry 4.0*, 1st ed, V. Bali (Ed.), pp. 39–55, CRC Press Taylor & Francis Group, USA, 2021, https://doi.org/10.1201/9781003146711-3.
3. Tounsi, W., What is cyber threat intelligence and how is it evolving?, in: *Cyber-Vigilance and Digital Trust: Cyber Security in the Era of Cloud Computing and IoT*, W. Tounsi (Ed.), ISTE Ltd, Washington, NJ, USA, 2019.
4. Dutta, A. and Kant, S., An overview of cyber threat intelligence platform and role of artificial intelligence and machine learning, in: *Proceedings of the 16th International Conference, ICISS 2020*, Lecture Notes in Computer Science book series (LNCS), pp. 81–86, Jammu, India, December 2020.
5. Mador, Z., Keep the dark web close and your cyber security tighter. *Comput. Fraud Secur.*, 2021, 1, 6–8, 2021.
6. Queiroz, A.L. and Keegan, B., Challenges of using machine learning algorithms for cybersecurity: a study of threatclassification models applied to social media communication data, in: *Cyber Influence and Cognitive 'reats*, J.M. Vladlena Benson (Eds.), Academic Press, Elsevier Inc., Cambride, MA, USA, 2020.
7. Sharma, A., Singh, U.K. *et al.*, Security and Privacy aspect of Cyber-Physical Systems, in: *Cyber Physical system: Concept and application*, A. Baliyan (Ed.), CRC Press Taylor & Francis Group, USA, 2022.
8. Sharma, A. and Singh, U.K., Investigation of Cloud Computing Security Issues & Challenges. *3rd International Conference on Integrated Intelligent Computing Communication & Security (ICIIC), 2021*, 2021, doi: https://doi.org/10.2991/ahis.k.210913.055.
9. Sharma, A., Singh, U.K. *et al.*, An investigation of security risk & taxonomy of Cloud Computing environment. *2021 2nd International Conference on Smart Electronics and Communication (ICOSEC)*, pp. 1056–1063, 2021, doi: 10.1109/ICOSEC51865.2021.9591954.
10. Sharma, A., Singh, U.K. *et al.*, A Comparative analysis of security issues & vulnerabilities of leading Cloud Service Providers and in-house University Cloud

platform for hosting E-Educational applications. *IEEE Mysore Sub Section International Conference (MysuruCon)*, 2021, ISBN: 978-0-7381-4662-1.

11. Top 5 cloud security breaches, https://www.cybertalk.org/2022/04/26/top-5-cloud-security-breaches-and-lessons/.
12. Latest cloud security news, https://portswigger.net/daily-swig/cloud-security.
13. Internet Security Threat Report, Internet Report, vol. 21, APRIL 2016.
14. Kaur, R. and Singh, M., Automatic Evaluation and Signature Generation Technique for Thwarting ZeroDay Attacks. *Second International Conference, SNDS 2014*, India, March 13-14, 2014, pp. 298–309.
15. White Paper, —ZERO-DAY DANGER: A Survey of Zero-Day Attacks and What They Say About the Traditional Security Model, FireEye Security Raimagined, 2015.
16. Diro, A.A. and Chilamkurti, N., Distributed attack detection scheme using deep learning approach for Internet of Things. *Future Gener. Comput. Syst.*, 82, 761–768, 2018, ISSN 0167-739X, https://doi.org/10.1016/j.future.2017.08.043.
17. Hegde, J. and Rokseth, B., Applications of machine learning methods for engineering risk assessment – A review. *Saf. Sci.*, 122, 104492, 2020, ISSN 0925-7535, https://doi.org/10.1016/j.ssci.2019.09.015.
18. Tan, L., Yu, K., Ming, F., Cheng, X., Srivastava, G., Secure and Resilient Artificial Intelligence of Things: A HoneyNet Approach for Threat Detection and Situational Awareness. *IEEE Consum. Electron. Mag.*, 11, 3, 69–78, 1 May 2022, doi: 10.1109/MCE.2021.3081874.
19. Lin, S.-S., Shen, S.-L., Zhou, A., Xu, Y.-S., Risk assessment and management of excavation system based on fuzzy set theory and machine learning methods. *Autom. Constr.*, 122, 103490, 2021, ISSN 0926-5805, https://doi.org/10.1016/j.autcon.2020.103490.
20. Mishra, A. K., Tyagi, A. K., Dananjayan, S., Rajavat, A., Rawat, H., Rawat, A., Revolutionizing Government Operations: The Impact of Artificial Intelligence in Public Administration, in: *Conversational Artificial Intelligence*, pp. 607–634, 2024.
21. Sharma, A. and Singh, U.K., Deployment model of e-educational cloud for departmental academics automation using open source. *HTL J.*, 27, 5, 36, 2021, ISSN 1006-6748. https://doi.org/10.37896/HTL27.5/3535.
22. Olzak, T., Mitigating Zero Day Attacks with a Detection, Prevention and Response Strategy, 2021, [online], Available: https://www.toolbox.com/it-security/vulnerability-management/articles/mitigating-zero-day-attacks/. [Accessed: 16-03-2022].
23. Nahar, S., Pithawa, D., Bhardwaj, V., Rawat, R., Rawat, A., Pachlasiya, K., Quantum technology for military applications, in: *Quantum Computing in Cybersecurity*, pp. 313–334, 2023.
24. Narayan, V. and Shaju, B., Malware and Anomaly Detection Using Machine Learning and Deep Learning Methods, in: *Handbook of Research on Machine*

and Deep Learning Applications for Cyber Security, 2020, DOI: 10.4018/978-1-5225-9611-0.ch006.

25. Namdev, A., Patni, D., Dhaliwal, B. K., Parihar, S., Telang, S., Rawat, A., Potential Threats and Ethical Risks of Quantum Computing, in: *Quantum Computing in Cybersecurity*, pp. 335–352, 2023.
26. Shrivastava, M., Patil, R., Bhardwaj, V., Rawat, R., Telang, S., Rawat, A., Quantum Computing and Security Aspects of Attention-Based Visual Question Answering with Long Short-Term Memory, in: *Quantum Computing in Cybersecurity*, pp. 395–412, 2023.
27. Sikarwar, R., Shakya, H. K., Kumar, A., Rawat, A., Advanced Security Solutions for Conversational AI, in: *Conversational Artificial Intelligence*, pp. 287–301, 2024.
28. Ahmad, R., Alsmadi, I., Alhamdani, W. *et al.*, Zero-day attack detection: a systematic literature review. *Artif. Intell. Rev.*, **56**, 10733–10811, 2023, https://doi.org/10.1007/s10462-023-10437-z.
29. Chang, V., Golightly, L., Modesti, P., Xu, Q.A., Doan, L.M.T., Hall, K., Boddu, S., Kobusińska, A., A Survey on Intrusion Detection Systems for Fog and Cloud Computing. *Future Internet*, 14, 3, 89, 2022, https://doi.org/10.3390/fi14030089.
30. Yasmeen, K. and Adnan, M., Zero-day and zero-click attacks on digital banking: a comprehensive review of double trouble. *Risk Manage.*, 25, 25, 2023, https://doi.org/10.1057/s41283-023-00130-4.
31. Nair, D. and Mhavan, N., Augmenting Cybersecurity: A Survey of Intrusion Detection Systems in Combating Zero-day Vulnerabilities, in: *Smart Analytics, Artificial Intelligence and Sustainable Performance Management in a Global Digitalised Economy (Contemporary Studies in Economic and Financial Analysis, vol. 110A)*, P. Tyagi, S. Grima, K. Sood, B. Balamurugan, E. Özen, T. Eleftherios, (Eds.), pp. 129–153, Emerald Publishing Limited, Leeds, 2023, https://doi.org/10.1108/S1569-37592023000110A007.
32. Sepúlveda Estay, D.A., A system dynamics, epidemiological approach for high-level cyber-resilience to zero-day vulnerabilities. *J. Simul.*, *17*, 1, 1–16, 2023, https://doi.org/10.1080/17477778.2021.1890533.
33. Nkongolo, M. and Tokmak, M., Zero-Day Threats Detection for Critical Infrastructures, in: *South African Institute of Computer Scientists and Information Technologists.* SAICSIT 2023. Communications in Computer and Information Science, vol. 1878, A. Gerber, and M. Coetzee, (Eds.), Springer, Cham, 2023, https://doi.org/10.1007/978-3-031-39652-6_3.
34. Singh, C. and Jain, A.K., A comprehensive survey on DDoS attacks detection & mitigation in SDN-IoT network. *e-Prime Adv. Electr. Eng. Electron. Energy*, 8, 100543, 2024, ISSN 2772-6711, https://doi.org/10.1016/j.prime.2024.100543.
35. Noonia, A., Beg, R., Patidar, A., Bawaskar, B., Sharma, S., Rawat, H., Chatbot vs Intelligent Virtual Assistance (IVA), in: *Conversational Artificial Intelligence*, pp. 655–673, 2024.

36. Ilca, L.F., Lucian, O.P., Balan, T.C., Enhancing Cyber-Resilience for Small and Medium-Sized Organizations with Prescriptive Malware Analysis, Detection and Response. *Sensors*, 23, 15, 6757, 2023, https://doi.org/10.3390/s23156757.
37. Silvia Priscila, S., Sharma, A., Vanithamani, S., Ahmad, F., Mahaveerakannan, R., Alrubaie, A.J., Jagota, V., Singh, B.K., [Retracted] Risk-Based Access Control Mechanism for Internet of Vehicles Using Artificial Intelligence. *Secur. Commun. Netw.*, 2022, Article ID 3379843, 13, 2022, https://doi.org/10.1155/2022/3379843.
38. Zekrifa, D.M.S., Sharma, A., Satyam, Patankar, A.J., Bamane, K.D., Data Analyzing with Cloud Computing Including Related Tools and Techniques. *Int. J. Intell. Syst. Appl. Eng.*, 11, 9s, 233–238, 2023, Retrieved from https://ijisae.org/index.php/IJISAE/article/view/3112.
39. Sharma, A. and Dixit, V., A Survey on Reliability of Cloud Applications. *Int'l Conf. Emerging Trends and Developments in Science, Management and Technology (ICETDSMT '13)*, pp. 476–481, 2013/3, ISBN: 978-81-924342-2-3.
40. Zekrifa, D., Sharma, A., Sharma, D., Sharma, R., Rai, S., Pillai, A., A System Based on AI and M1 Enhanced to Investigate Physiological Markers for User Forecasting Decision-Making, pp. 2487–2490, 2023, 10.1109/IC3I59117.2023.10398037.
41. Vijayakumar, T., Ramalakshmi, K., Priyadharsini, C., Vasanthakumar, S., Shaina, Sharma, A., Bio-Inspired Optimization Algorithm on Cloud based Image Retrieval System using Deep Features. *2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, Trichy, India, pp. 871–876, 2022, doi: 10.1109/ICAISS55157.2022.10010739.
42. Pithawa, D., Nahar, S., Bhardwaj, V., Rawat, R., Dronawat, R., Rawat, A., Quantum Computing Technological Design Along with Its Dark Side, in: *Quantum Computing in Cybersecurity*, pp. 295–312, 2023.
43. Carrier, T., Victor, P., Tekeoglu, A., Lashkari, A.H., Detecting Obfuscated Malware using Memory Feature Engineering. *The 8th International Conference on Information Systems Security and Privacy (ICISSP)*, 2022.
44. Mahdavifar, A.S., Maleki, N., Lashkari, A.H., Broda, M., Razavi, A.H., Classifying Malicious Domains using DNS Traffic Analysis. *The 19th IEEE International Conference on Dependable, Autonomic, and Secure Computing (DASC)*, Calgary, Canada, Oct. 25-28, 2021.
45. Apriorit, Zero-day Attacks Detection and Prevention Methods, 2022, [online]. Available: https://www.apriorit.com/dev-blog/450-zero-day-attack-detection. [Accessed: 20-4-2022].
46. Lashkari, A.H., Kaur, G., Rahali, A., DIDarknet: A Contemporary Approach to Detect and Characterize the Darknet Traffic using Deep Image Learning (https://dl.acm.org/doi/10.1145/3442520.3442521). *10th International Conference on Communication and Network Security*, Tokyo, Japan, November 2020.
47. Sharafaldin, I., Lashkari, A.H., Hakak, S., Ghorbani, A.A., Developing Realistic Distributed Denial of Service (DDoS) Attack Dataset and

Taxonomy (https://ieeexplore.ieee.org/abstract/document/8888419). *IEEE 53rd International Carnahan Conference on Security Technology*, Chennai, India, 2019.

48. Sharafaldin, I., Lashkari, A.H., Ghorbani, A.A., Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. *4th International Conference on Information Systems Security and Privacy (ICISSP)*, Portugal, January 2018.
49. Aldribi, A., Traore, I., Moa, B., Data Sources and Datasets for Cloud Intrusion Detection Modeling and Evaluation, in: *Cloud Computing for Optimization: Foundations, Applications, and Challenges*. Studies in Big Data, vol. 39, B. Mishra, H. Das, S. Dehuri, A. Jagadev (Eds.), pp. 333–366, Springer, USA, 2018.
50. Aldribi, A., Traore, I., Moa, B., Nwamuo, O., Hypervisor-Based Cloud Intrusion Detection through Online Multivariate Statistical Change Tracking, in: *Computers & Security*, vol. 88, Elsevier, USA, 2023, January 2020.
51. Rathore, N. and Rajavat, A., Smart Farming Based on IOT-Edge Computing: Applying Machine Learning Models For Disease And Irrigation Water Requirement Prediction In Potato Crop Using Containerized Microservices, in: *Precision Agriculture for Sustainability*, pp. 399–424, Apple Academic Press, USA, 2024.
52. Patsariya, M. and Rajavat, A., A Progressive Design of MANET Security Protocol for Reliable and Secure Communication. *Int. J. Intell. Syst. Appl. Eng.*, *12*, 9s, 190–204, 2024.
53. Rathi, M. and Rajavat, A., Investigations and Design of Privacy-Preserving Data Mining Technique for Secure Data Publishing. *Int. J. Intell. Syst. Appl. Eng.*, *11*, 9s, 351–367, 2023.
54. Dubey, P. and Rajavat, A., Effective K-means clustering algorithm for efficient data mining, in: *2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN)*, pp. 1–6, IEEE, 2023, May.
55. Chauhan, D., Singh, C., Rawat, R., Chouhan, M., Conversational AI Applications in Ed-Tech Industry: An Analysis of Its Impact and Potential in Education, in: *Conversational Artificial Intelligence*, pp. 411–433, 2024.
56. Chauhan, D., Singh, C., Rawat, R., Dhawan, M., Evaluating the Performance of Conversational AI Tools: A Comparative Analysis, in: *Conversational Artificial Intelligence*, pp. 385–409, 2024.

26

Predicting Rumors Spread Using Textual and Social Context in Propagation Graph with Graph Neural Network

Siddharath Kumar Arjaria[1], Hardik Sachan[1], Satyam Dubey[1], Ayush Pandey[1], Mansi Gautam[1], Nikita Gupta[1] and Abhishek Singh Rathore[2]*

[1]Department of Information Technology, Rajkiya Engineering College, Banda (U.P.), India
[2]Department of Computer Science & Engineering, Shri Vaishnav Vidyapeeth Vishwavidyalaya, Indore, India

Abstract

In today's age of digital advancement, it has become very easy for people to share their opinions on the internet. Social media platform like Twitter, Facebook, etc. connects people to share the latest news in the world. Rumor is a term used to describe unsubstantiated information or rumors that are widely shared by users but not verified by reliable sources. With this freedom of sharing tweets and opinions, it becomes very easy to spread a rumor. Rumors can create chaos in the public domain that can trigger fear, damage a person's image, and can also affect other sectors like the economy, etc. To control the damage, a strong rumor detection system is needed so that it can detect rumors with better accuracy. Previous models only consider the tweep profile and authenticity of the profile while this paper aims to develop a rumor detection system that uses the previous history of users, propagating pattern of tweets, and semantic relation of text content. At last, attention mechanism is also used to enhance the accuracy of the model. This comprehensive methodology harnesses the strengths of graph-based and textual analysis for nuanced and accurate rumor detection model on social media, particularly Twitter.

***Keywords*:** Rumors detection, graph neural network, social media network, twitter, bert, word2vec, representation learning

**Corresponding author*: abhishekatujjain@gmail.com

Rajesh Kumar Chakrawarti, Ranjana Sikarwar, Sanjaya Kumar Sarangi, Samson Arun Raj Albert Raj, Shweta Gupta, Krishnan Sakthidasan Sankaran and Romil Rawat (eds.) Natural Language Processing for Software Engineering, (389–402) © 2025 Scrivener Publishing LLC

26.1 Introduction

In today's digital age, where social media like Twitter and Facebook have become integral parts of our daily lives, the dependency on these platforms to disseminate information has skyrocketed. However, this increased reliance also brings with it the risk of the rapid spread of rumors, capable of causing significant disruptions in individuals' lives. The pervasive nature of social media makes it a fertile ground for the propagation of misinformation, necessitating a critical examination of the impact rumors can have on society.

The potential for rumors to bring about massive changes in the lives of individuals cannot be understated. False information, once unleashed on social media, can quickly gain traction, leading to widespread panic, confusion, and damage to reputations. Recognizing the urgency of the situation, the need for a robust rumor detection system has become increasingly apparent. As rumors can have far-reaching consequences, from affecting personal relationships to influencing political climates, the development of effective detection mechanisms has become a pressing concern.

Mitigating the effects of rumors is not only essential for safeguarding individual integrity but also for maintaining the stability of communities and societies at large. The destructive potential of false information necessitates proactive measures to identify and counteract rumors before they can cause irreparable harm. As a result, the focus on rumor detection has intensified in recent years, with researchers and technologists exploring various techniques to address this burgeoning issue.

The proposed methodology integrates multiple stages to form a comprehensive framework for rumor detection. Firstly, a hybrid data collection strategy combining API interfaces and third-party crawler programs is employed to extract Twitter rumors information from the Twitter public page. Feature extraction encompasses both content and context analysis, incorporating textual content and social context features of users involved in the tweet, such as followers, interests, and verified accounts.

Textual embeddings are generated using advanced models like Word2Vec [1] and Bidirectional Encoder Representations from Transformers (BERT) [2], enabling the system to understand the semantic relationships within tweet text. A unique contribution lies in the construction of a propagation graph, where the tweet serves as the root node and involved users as leaf nodes. Graph Neural Networks (GNNs) [3] are then applied to extract features capturing the influence and engagement of users. This involves message passing, node updating, and iterative refinement processes, allowing complex relationships within the network to be captured effectively.

Furthermore, an attention-based mechanism [4] combines textual embeddings with GNN-derived user involvement embeddings to obtain a precise and less redundant tweet representation. Finally, a neural classifier is trained on the integrated embeddings to predict the probability of a tweet being a rumor. This comprehensive methodology offers an advanced and adaptive solution, leveraging the strengths of both graph-based and textual analysis for more accurate and nuanced rumor detection on social media platforms, particularly Twitter.

The paper is organized as follows. Section 26.2 presents a review of the literature done for rumor identification. Section 26.3 discusses the proposed methodology, followed by section 26.4. Section 26.4 discusses the results of the proposed work.

26.2 Literature Review

This section covers the work done by different authors to detect rumors using machine learning techniques. Despite extensive research from various viewpoints, the dissemination mechanism of rumors has not been properly examined. In terms of monitoring, it is difficult to concretely define propagation patterns.

Ma *et al.* [5] applied recurrent neural networks to extract rumors sequentially. Yang *et al.* [6] extracted location and client-based [7] features to identify the rumors.

Kipf and Welling [8] presented a semi-supervised approach to graphical data. First-order approximations are applied on convolutional neural networks (CNN) for layer-wise propagation. Li *et al.* [9] added gate recurrent units [10] to update the output sequence of GNNs [3]. Xu *et al.* [11] combined GNNs with hierarchical analysis of text and propagation structure. The text-granularity and word-granularity features were integrated to provide hierarchically fused features that reflect word nodes in source text graphs.

Chen *et al.* [12] proposed a soft attention-based recurrent neural network that focuses on selective attributes and learns hidden representations of social posts over time. Gao *et al.* [13] used uncertainty semantics with event context to generate rumor representation using attention networks. Kwon *et al.* [14] identified three feature sets. Temporal and structural feature sets are extracted through time series which are later combined with linguistic feature sets to identify rumors in comparison with standard features. Later Kwon *et al.* [15] identified rumors spreading patterns using a three to two months window. Constraints have been proposed to remove options for using subject information as features, to make preconceived notions about the nature of the network, or to make preliminary assumptions about user characteristics about the information signal [16].

Song *et al.* [17] identified reposting sequences on microblogs to find detection points for early rumors detection. A convolutional neural network is proposed to identify the rumors spreading behavior of a person focusing on the past posts of the users, his reactions, and reposting the rumors [18]. Wang *et al.* [19] used a partition-based approach to identify rumors propagation structure and learn through a temporal attention network. The content and temporal structures are then fused to predict rumors in the network. Zhang and Hara [20] proposed a probabilistic model to classify malicious users, as the source of rumors. Rumors traceability under incomplete information is a real issue. A symmetric difference between the contaminated set at the final layer of the social network and the recently tainted set is identified to trace the sources of rumors [21].

Wei *et al.* [22] proposed a transformer-based learning model to capture deep rumor sequences and their semantic relationships. Zhu *et al.* [23] studied rumor propagation by analyzing the behavior of users through the reaction-diffusion dynamic model. Wang *et al.* [24] proposed an unsupervised model based on neighborhood information transmission. A generation-based improvement strategy is used to improve the results of the unsupervised model.

Patel *et al.* [25] tackle the class imbalance challenge in rumor detection through a combination of contextualized data augmentation, and novel GNN models. The oversampling technique creates synthetic examples for underrepresented classes in the dataset by utilizing contextualized data augmentation. The basic notion uses a thread's tweet selection for augmentation, which is made feasible by giving a non-random selection criterion that directs the augmentation process to relevant tweets.

Huang *et al.* [26] developed a tweet-word-user heterogeneous graph based on the text content and source tweet propagation of rumors. It captures the global semantic relations of text contents, as well as the global structural information of source tweet propagations, to identify rumors. Han *et al.* [27] suggested a classifier that provides balanced performance on both current and new datasets by employing continuous learning approaches to gradually train GNNs.

Wu *et al.* [28], created a gated propagation network based on the structure of Twitter postings. The propagation network constantly adjusts the weights of each node by integrating attention methods for improved classification tasks. Ali and Malik [29] employed Word2Vec and BERT to add contextual information in rumor identification.

Furthermore, past research employed emotional and conversational characteristics to identify rumors without taking into account distinct emotions, as well as effective PoS elements in content-based techniques. Similarly, the bulk of previous studies employed content-based techniques to feature development, while contemporary context-based approaches were not investigated. In addition, semantic and social characteristics are also not fully utilized for rumor identification.

26.3 Proposed Methodology

The proposed framework, TTRD (Tweep Tendency-aware Rumors News Detection), tackles the challenge of identifying rumors within news content. TTRD operates in three key stages:

a. Unveiling Tweep Tendencies: Analyze the historical social media activity (e.g., tweets) of tweeps who interacted with a news item to uncover their preexisting inclinations and tendencies. Text representation techniques like word2vec and BERT process their past posts to glean these insights. The news textual content is similarly encoded.
b. Capturing Network Dynamics: Moving beyond individual tendencies, Harness the "social fingerprint" of the news by constructing a network reflecting its propagation on social media platforms (e.g., retweet relationships on Twitter). This network analysis captures the broader context surrounding the news and its resonance with different tweep groups.
c. Integrating Tweep and Network Signals: The final stage seamlessly combines these insights through a multi-layered information fusion process. Leverage Graph Neural Networks (GNNs) to analyze the social network and extract an "involvement embedding" reflecting tweep interaction within it. This embedding is then combined with the encoded textual content of the news to create a comprehensive representation. Subsequent sections delve deeper into each component, detailing how to extract tweep tendencies, capture network dynamics, and ultimately fuse these elements by applying attention mechanisms to identify rumors within news content.

The proposed framework aims to detect tendencies in tweep-generated rumors news. By analyzing the news content alongside the interactions of active users on social media platforms, we construct a news propagation graph to capture the broader context. The intrinsic details are gleaned from tweeps' historical posts and the content of the news articles. These diverse sources of information are integrated using a GNN encoder. The resulting news representation combines user engagement features with textual attributes, which are then inputted into a neural classifier to assess the credibility of the news piece as shown in Figure 26.1: The proposed TTRD framework.

26.3.1 Tweep Tendency Encoding

Accurately modeling tweep tendencies solely from their social network data poses a significant challenge. Following prior work that infers personality, sentiment, and stance based on historical posts, propose an implicit

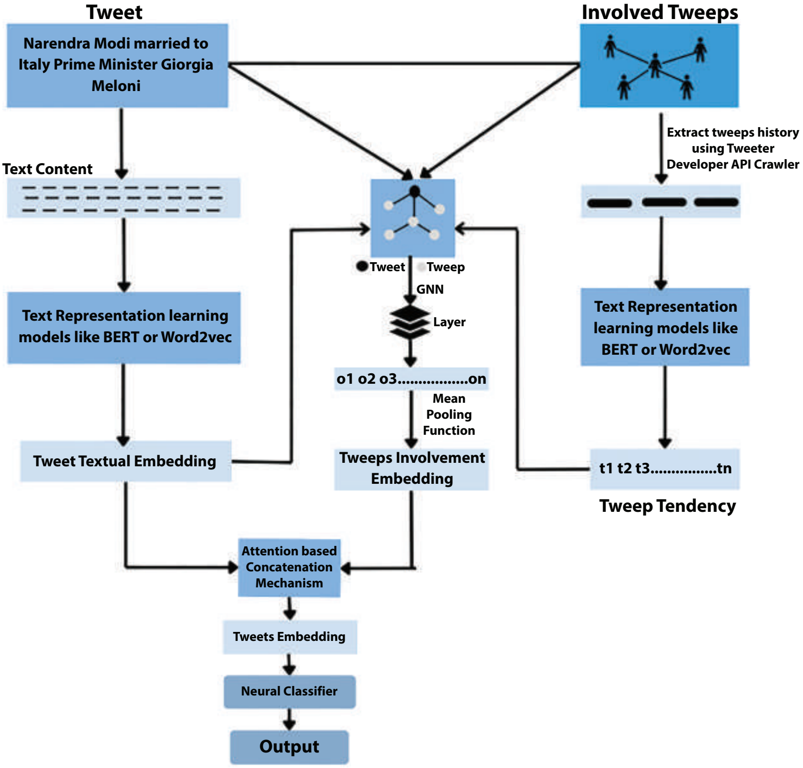

Tweet
Narendra Modi married to Italy Prime Minister Giorgia Meloni
Involved Tweeps
Text Content
Extract tweeps history using Tweeter Developer API Crawler
● Tweet
● Tweep
GNN
Layer
Text Representation learning models like BERT or Word2vec
Text Representation learning models like BERT or Word2vec
o1 o2 o3.................on
Mean Pooling Function
Tweet Textual Embedding
Tweeps Involvement Embedding
t1 t2 t3................tn
Tweep Tendency
Attention based Concatenation Mechanism
Tweets Embedding
Neural Classifier
Output

Figure 26.1 The proposed TTRD framework.

approach leveraging tweet analysis. However, existing rumors news datasets lack such tendency information.

To address this gap, the researchers utilize the Rumours News Net dataset, which provides news content and corresponding Twitter involvement data. By harnessing the Twitter Developer API, they retrieve the most recent 200 tweets (approximately 20 million in total) from accounts that retweeted each news item. To account for inaccessible accounts (e.g., suspended or deleted), they substitute random tweets with tweeps sharing the same news, ensuring the integrity of the news propagation cascade for effective context analysis.

Before using text representation learning methods, the data are preprocessed by removing special characters like "@" mentions and URLs. To capture both news content and tweep tendencies, two pre-trained language models are employed, and we follow these:

a. Word2vec: Leverage spaCy's pre-trained vectors for 680k words to encode semantic relationships between words and respective sentences. For each tweep, obtain their tendency representation by averaging the vectors of existing words in their combined 200 recent tweets. Similarly, create a news text embedding following the same process.
b. BERT: Use the cased BERT-Large model to encode news and tweep information as sequences. While the news content can be directly fed into BERT (max length 512 tokens), individual tweet encoding is required due to length limitations. Encode each tweet separately and average their resulting vectors to represent a tweep's tendency. To optimize encoding speed for shorter tweets compared to news text, empirically set the maximum tweet length to 16 tokens.

Combining these pre-trained models and carefully handling inaccessible accounts, creates a robust approach to encode tweep tendencies implicitly from their tweet history. This paves the way for further analysis and applications aimed at detecting rumors within news content.

26.3.2 Network Dynamics Extraction

The social fingerprint of a news piece on social media is influenced by the collective interactions of all users engaging with it. One has to delve into this fingerprint by constructing a news propagation graph. He/She has to imagine this graph like a branching tree (refer to Figure 26.1): the root represents the news, and each branch signifies a tweep who shared it. This paper focuses on rumor detection on Twitter, using this approach as a proof of concept.

To build these Twitter networks, one has to follow established methods. He/She has to consider a news item as node v_1 and tweeps who retweeted it as nodes v_2 to v_n, listed chronologically. Two rules guide how to establish the news propagation path:

Rule 1: Following the Trend: If tweep v_i retweets the news after any previous retweeter in the sequence (v_1 to v_n), assume the news spread from the later tweep to v_i. This logic stems from the assumption that recent tweets are more likely to be seen and retweeted.
Rule 2: Leveraging Influence: If v_i doesn't follow any previous retweeters, conservatively assume the news spread from the tweeps with the highest follower count. This reflects the higher visibility of tweets from these tweeps based on Twitter's content distribution mechanisms.

Utilize these guidelines to create news dissemination graphs on Twitter. This method can also be expanded to other social networking sites [33–35] such as Facebook [29–32].

26.3.3 Extracted Information Integration

Previous research has shown that combining tweep features with news propagation graphs can significantly enhance fake news detection accuracy. Graph Neural Networks (GNNs) excel at this task due to their ability to seamlessly integrate node features and graph structure within a unified learning framework. Propose a hierarchical information fusion approach that leverages this strength, further refined by an attention mechanism for a more precise and robust representation.

Stage 1: GNN-based tweep tendency and network dynamics
Information Integration:

1. Node Feature Preparation: Represent the news content using its textual embedding and each tweep's tendency using their embedding derived from historical tweets. These embeddings serve as node features in the news propagation graph.
2. GNN Feature Aggregation: Employing a GNN, aggregate the features of a node's neighbors to refine its representation.

This process captures the collective influence of related tweeps and the underlying network structure.

3. Pooling Function: Similar to graph classification models using GNNs, apply a pooling function (e.g., mean pooling) overall node embeddings to obtain a single embedding for the entire news propagation graph, effectively summarizing the collective "tweep involvement embedding."

Stage 2: Attention-based Refinement of Tweep Involvement Embedding: To further refine the tweep involvement embedding obtained from the GNN, introduce an attention mechanism that focuses on the most relevant aspects of each tweep's involvement:

1. Dot Product Attention: Employ a dot-product attention mechanism that calculates the similarity between the tweep involvement embedding and each tweep's embedding, considering various factors:
   - Tweep Tendency: Accounts with tendencies similar to the news content are assigned higher weights, reflecting their potential bias or alignment.
   - Tweep Profile Features: Verified accounts with more followers typically indicate greater credibility and receive higher weights.
   - Tweep Belief: Tweeps who express positive sentiments or beliefs aligned with the news content gain higher weights, reflecting potential agreement or support.
2. Weighted Combination: Based on these attention weights, create a weighted combination of the tweep involvement embedding and each tweep's embedding, highlighting the most relevant information for this particular news item.

Stage 3: Final News Embedding and Classification:

1. Concatenation: Enrich the news embedding by concatenating the refined tweep involvement embedding with the original news textual embedding. This combined representation captures both the content itself and the surrounding social context.
2. Multilayer Perceptron (MLP): The fused news embedding is fed into a two-layer MLP with two output neurons, predicting the probabilities of the news being fake or real.

3. Training and Optimization: The model is trained using a binary cross-entropy loss function and optimized with stochastic gradient descent (SGD).

26.4 Results and Discussion

Dataset: To explore the dissemination of rumors on social media, we have selected the Rumors dataset, which comprises information on both rumor and non-rumor tweets. This dataset includes data from two fact-checking websites as well as the corresponding social engagement on Twitter. By analyzing the tweet tendency and the involvement of Twitter users, we aim to gain insights into the propagation of rumors as shown in Table 26.1: Dataset and graph statistics.

For extracting the tweep information use an API crawler that extracts information of all the tweeps which are related to the tweet. Upon evaluation, observe that the TTRD model has a better performance compared to all other models. The experimental results of TTRD show that Social Context significantly improves the result, and the Tweet Tendency and attention mechanism could help when the text content of the tweet is limited. The attention mechanism gives importance to only those attributes that play a considerable role in spreading the rumors as given in Table 26.2: Performance Evaluation.

Second, all other models encode the text content (tweet) without considering the tweep history, which can tell that leveraging the tweep history could improve the Rumor detection performance. The TTRD with the best performance on both datasets uses BERT as the text encoder as shown in Table 26.3: Rumor detection model performance with different node features models and types.

Table 26.1 Dataset and graph statistics.

Dataset	#Graphs (#Rumour)	#Total Nodes	#Total Edge	#Avg. Nodes per Graph
PolitiFact (POL)	360 (148)	41,054	42,740	135
Gossip cop (GOS)	5556 (2830)	314,262	305,457	62

Table 26.2 Performance evaluation.

	Model	POL		GOS	
		ACC	F1	ACC	F1
Tweets Only	BERT+MLP	72.06	72.03	85.05	85.35
	word2vec+MLP	77.47	76.86	85.61	85.9
Tweets+ Tweeps Involved	GNN-CL	63.90	63.25	95.28	95.34
	GCNFN	84.16	84.56	96.58	96.46
	TTRD (ours)	84.82%	84.85%	97.23%	97.22%

Table 26.3 Rumor detection model performance with different node features models and types.

Feature	POL				GOS			
	Graph SAGE		GCNFN		Graph SAGE		GCNFN	
	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1
ProRile	77.58	77.42	77.04	76.92	92.69	92.56	89.10	89.06
word2vec	80.64	80.71	80.74	80.81	96.91	96.89	95.07	95.05
BERT	84.82	84.73	83.86	83.54	97.43	97.22	96.18	96.17

26.5 Conclusion

Rumors can spread quickly and unpredictably on social media platforms, making automatic rumor detection technology a necessity. In this paper, we developed a deep attention-based rumor detection model. The model is divided into 3 stages. Initially, news contents and tweeps are extracted for news propagation network. Tweeps are used to identify social fingerprints. Then a propagation network is constructed. The graph neural network is used to fuse the news contents and graphical structure. It provides the unified learning framework. An attention-based refinement is applied to GNN to find news embedding.

At present, images and videos are major sources of rumors and fake news. And such contents increase when a country faces national elections. Future work of this article is to extract features that are suspicious and integrate them to our proposed propagation network.

References

1. Mishra, A. K., Tyagi, A. K., Dananjayan, S., Rajavat, A., Rawat, H., Rawat, A., Revolutionizing Government Operations: The Impact of Artificial Intelligence in Public Administration, in: *Conversational Artificial Intelligence*, pp. 607–634, 2024
2. Nahar, S., Pithawa, D., Bhardwaj, V., Rawat, R., Rawat, A., Pachlasiya, K., Quantum technology for military applications, in: *Quantum Computing in Cybersecurity*, pp. 313–334, 2023.
3. Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G., The Graph Neural Network Model. *IEEE Trans. Neural Networks*, 20, 1, 61–80, Jan. 2009, doi: 10.1109/TNN.2008.2005605.
4. Bahdanau, D., Cho, K., Bengio, Y., Neural Machine Translation by Jointly Learning to Align and Translate. Sep. 2014, [Online]. Available: http://arxiv.org/abs/1409.0473.
5. Ma, J. *et al.*, Detecting rumours from microblogs with recurrent neural networks, in: *IJCAI International Joint Conference on Artificial Intelligence*, vol. 2016-January, pp. 3818–3824, 2016.
6. Yang, F., Liu, Y., Yu, X., Yang, M., Automatic detection of rumour on Sina Weibo, in: *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, Aug. 2012, pp. 1–7, doi: 10.1145/2350190.2350203.
7. Rawat, R., Telang, S., William, P., Kaur, U., CU, O. K. (Eds.)., *Dark Web Pattern Recognition and Crime Analysis Using Machine Intelligence*. IGI Global, 2022.
8. Sikarwar, R., Shakya, H.K., Kumar, A., Rawat, A., Advanced Security Solutions for Conversational AI, in: *Conversational Artificial Intelligence*, pp. 287–301, 2024.
9. Pithawa, D., Nahar, S., Bhardwaj, V., Rawat, R., Dronawat, R., Rawat, A., Quantum Computing Technological Design Along with Its Dark Side, in: *Quantum Computing in Cybersecurity*, pp. 295–312, 2023.
10. Namdev, A., Patni, D., Dhaliwal, B.K., Parihar, S., Telang, S., Rawat, A., Potential Threats and Ethical Risks of Quantum Computing, in: *Quantum Computing in Cybersecurity*, pp. 335–352, 2023.
11. Xu, S. *et al.*, Rumour detection on social media using hierarchically aggregated feature via graph neural networks. *Appl. Intell.*, 53, 3, 3136–3149, Feb. 2023, doi: 10.1007/s10489-022-03592-3.
12. Shrivastava, M., Patil, R., Bhardwaj, V., Rawat, R., Telang, S., Rawat, A., Quantum Computing and Security Aspects of Attention-Based Visual Question Answering with Long Short-Term Memory, in: *Quantum Computing in Cybersecurity*, pp. 395–412, 2023
13. Gao, Y., Han, X., Li, B., A Neural Rumour Detection Framework by Incorporating Uncertainty Attention on Social Media Texts, in: *Cognitive Computing – ICCC 2019*, pp. 91–101, 2019.

14. Noonia, A., Beg, R., Patidar, A., Bawaskar, B., Sharma, S., Rawat, H., Chatbot vs Intelligent Virtual Assistance (IVA), in: *Conversational Artificial Intelligence*, pp. 655–673, 2024
15. Rawat, R., Kaur, U., Khan, S. P., Sikarwar, R., Sankaran, K. (Eds.), *Using Computational Intelligence for the Dark Web and Illicit Behavior Detection*. IGI Global, 2022, https://doi.org/10.4018/978-1-6684-6444-1
16. Rawat, R., Telang, S., William, P., Kaur, U., C.U., O. (Eds.), *Dark Web Pattern Recognition and Crime Analysis Using Machine Intelligence*. IGI Global, 2022, https://doi.org/10.4018/978-1-6684-3942-5
17. Song, C., Yang, C., Chen, H., Tu, C., Liu, Z., Sun, M., CED: Credible Early Detection of Social Media Rumours. *IEEE Trans. Knowl. Data Eng.*, 33, 8, 3035–3047, Aug. 2021, doi: 10.1109/TKDE.2019.2961675.
18. Rawat, R., Chakrawarti, R.K., Sarangi, S.K., Patel, J., Bhardwaj, V., Rawat, A., Rawat, H. (eds.), in: *Quantum Computing in Cybersecurity*. John Wiley & Sons, 2023. https://onlinelibrary.wiley.com/doi/book/10.1002/9781394167401
19. Suthar, H., Rawat, H., Gayathri, M., Chidambarathanu, K., Techno-Nationalism and Techno-Globalization: A Perspective from the National Security Act, in: *Quantum Computing in Cybersecurity*, pp. 137–164, 2023
20. Zhang, Y. and Hara, T., A probabilistic model for malicious user and rumour detection on social media, in: *Proceedings of the Annual Hawaii International Conference on System Sciences*, vol. 2020-January, 2020, doi: 10.24251/hicss.2020.309.
21. Li, Y., Zhu, J., Wang, G., Huang, J., Rumour source identification under incomplete information in dynamic social network. *Xitong Gongcheng Lilun yu Shijian/Syst. Eng. Theory Pract.*, 43, 4, 1132–1144, Apr. 2023, doi: 10.12011/SETP2022-0262.
22. Singh, R., and Rajavat, A., Short-Lived Social Vehicular Network, in: *2024 International Conference on Electronics, Computing, Communication and Control Technology (ICECCC)*, pp. 1–9, IEEE, 2024, May
23. Singh, R., and Rajavat, A., VANET Security Using Blockchain, in: *Fostering Multidisciplinary Research for Sustainability*, p. 159, 2024
24. Vyas, G., Vyas, P., Muzumdar, P., Chennamaneni, A., Rajavat, A., Rawat, R., Extracting and Analyzing Factors to Identify the Malicious Conversational AI Bots on Twitter, in: *Conversational Artificial Intelligence*, pp. 71–83, 2024
25. Rawat, R., and Rajavat, A., Perceptual Operating Systems for the Trade Associations of Cyber Criminals to Scrutinize Hazardous Content. *Int. J. Cyber Warf. Terror. (IJCWT),* 14, 1, 1–19, 2024.
26. Chauhan, D., Singh, C., Rawat, R., Dhawan, M., Evaluating the Performance of Conversational AI Tools: A Comparative Analysis, in: *Conversational Artificial Intelligence*, pp. 385–409, 2024.
27. Rawat, R., Chakrawarti, R.K., Sarangi, S.K., Choudhary, R., Gadwal, A.S., Bhardwaj, V. (eds.), *Robotic Process Automation*. John Wiley & Sons, 2023.
28. Manmadhan, S., and Kovoor, B.C., Visual question answering: a state-of-the-art review. *Artif. Intell. Rev.*, 53, 8, 5705–5745, 2020.

29. Ali, G. and Malik, M.S.I., Rumour identification on Twitter as a function of novel textual and language-context features. *Multimed. Tools Appl.*, 82, 5, 7017–7038, Feb. 2023, doi: 10.1007/s11042-022-13595-4.
30. Rathore, N. and Rajavat, A., Smart Farming Based on IOT-Edge Computing: Applying Machine Learning Models For Disease And Irrigation Water Requirement Prediction In Potato Crop Using Containerized Microservices, in: *Precision Agriculture for Sustainability*, pp. 399–424, Apple Academic Press, USA, 2024.
31. Patsariya, M. and Rajavat, A., A Progressive Design of MANET Security Protocol for Reliable and Secure Communication. *Int. J. Intell. Syst. Appl. Eng.*, *12*, 9s, 190–204, 2024.
32. Rathi, M. and Rajavat, A., Investigations and Design of Privacy-Preserving Data Mining Technique for Secure Data Publishing. *Int. J. Intell. Syst. Appl. Eng.*, *11*, 9s, 351–367, 2023.
33. Dubey, P. and Rajavat, A., Effective K-means clustering algorithm for efficient data mining, in: *2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN)*, pp. 1–6, IEEE, 2023, May.
34. Chauhan, D., Singh, C., Rawat, R., Chouhan, M., Conversational AI Applications in Ed-Tech Industry: An Analysis of Its Impact and Potential in Education, in: *Conversational Artificial Intelligence*, pp. 411–433, 2024.
35. Chauhan, D., Singh, C., Rawat, R., Dhawan, M., Evaluating the Performance of Conversational AI Tools: A Comparative Analysis, in: *Conversational Artificial Intelligence*, pp. 385–409, 2024.

27

Implications, Opportunities, and Challenges of Blockchain in Natural Language Processing

Neha Agrawal[1]*, Balwinder Kaur Dhaliwal[2], Shilpa Sharma[3], Neha Yadav[4] and Ranjana Sikarwar[5]

[1]*Department of Computer Science Engineering, Shri Vaishnav Vidyapeeth Vishwavidyalaya, Indore, India*
[2]*Department of Computer Science and Engineering, Lovely Professional University, Jalandhar, Punjab, India*
[3]*Master of Computer Applications Department, IPS Academy, Sanwer Campus, Indore, MP, India*
[4]*Institute of Engineering and Science, IPS Academy, Indore, MP, India*
[5]*Amity University, Gwalior, MP, India*

Abstract

Blockchain technology is becoming a game-changer in sectors including healthcare, banking, and supply chain management. Its possible influence on Natural Language Processing (NLP) is still mainly untapped, nevertheless. The goal of natural language processing (NLP), a branch of artificial intelligence, is to make text resources comprehensible, interpreted, and producible by computers. This article examines how blockchain and natural language processing might work together, emphasizing the possible advantages for enhancing data security, trust, and transparency. NLP is challenged by unstructured text data, which includes problems with data quality and origin verifiability. The blockchain's decentralized and unchangeable properties offer a remedy by fostering confidence and guaranteeing the authenticity of textual materials. The article explores the use of blockchain technology in copyright protection, plagiarism detection, and content verification, emphasizing its capacity to verify material and stop intellectual property theft. Furthermore, decentralized, and cooperative textual data annotation and categorization are made easier by the combination of blockchain technology and natural language processing, which raises

**Corresponding author*: agrawal.na@gmail.com

Rajesh Kumar Chakrawarti, Ranjana Sikarwar, Sanjaya Kumar Sarangi, Samson Arun Raj Albert Raj, Shweta Gupta, Krishnan Sakthidasan Sankaran and Romil Rawat (eds.) Natural Language Processing for Software Engineering, (403–424) © 2025 Scrivener Publishing LLC

the Caliber and accessibility of annotated datasets. But to properly utilize the potential of this integration, issues with scalability, privacy, and transparency need to be resolved. This paper advances the development of secure, dependable, and effective NLP systems through a thorough study of the body of existing literature, use case analysis, discussion of challenges, and suggestion of future research areas.

***Keywords*:** Natural language processing, blockchain, artificial intelligence, decentralization, security

27.1 Introduction

Blockchain technology is becoming more and more potent, revolutionizing a wide range of sectors, including supply chain management, healthcare, and finance. By combining with Blockchain technology, AI applications become more powerful and address issues with trust and data integrity. Blockchain technology and artificial intelligence together create a safe foundation for innovation in a time when industries are becoming more and more reliant on data-driven decision-making [1, 2]. Blockchain and AI work together to create a strong coalition that addresses problems with efficiency, transparency, and data security. As a result, the market for blockchain and AI technologies is expected to grow at a compound annual growth rate (CAGR) of 25.3% from 2020 to 2025, surpassing $703 million as shown in Figure 27.1 [3].

Natural language processing (NLP) is a subfield of artificial intelligence (AI) that aims to enable computer devices and applications to comprehend, interpret, and produce organized and accessible corpora in human

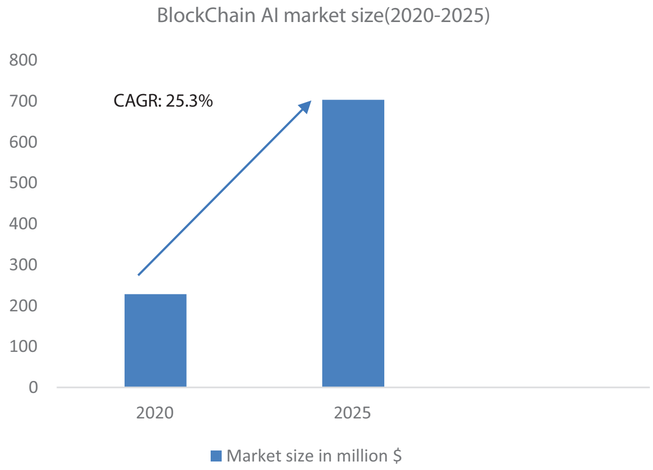


Figure 27.1 Blockchain AI market size.

languages [4]. The combination of blockchain technology and natural language processing has great promise for enhancing data security, fostering trust, and increasing transparency. Its possible influence on Natural Language Processing (NLP) is still largely unknown, though.

NLP researchers face many difficulties when dealing with unstructured textual data that may be found on the Internet, such as news stories, social media posts, and scientific publications. Data confidentiality, data quality, and data origin reliability are a few possible problems. By ensuring confidence and integrity of textual resources, blockchain technology—with its decentralized and immutable register—offers a solution. Blockchain technology can help NLP researchers and developers address important problems like secure data sharing, copyright protection, plagiarism detection, and locating data references.

Blockchain technology is being applied in NLP for the purposes of copyright protection and content verification, or authentication. This is achieved by keeping the original text's distinct hash values on the blockchain. This creates a decentralized, tamper-proof block that makes it possible to verify the text's integrity and authorship. Additionally, by comparing hashes of suspected material with those blocks on the blockchain, blockchain technology can help detect plagiarism [5] more effectively while guaranteeing authenticity and avoiding intellectual property (IP) infringement.

Additionally, the combination of blockchain technology and natural language processing creates new avenues for collaborative and decentralized textual data annotation and classification. Blockchain-based systems can improve the availability and quality of annotated datasets, which will help NLP research and model development, by incentivizing contributors through tokenization and intelligent categorization.

But despite its enormous potential, there are obstacles to overcome before blockchain technology can be applied to NLP. This involves worries about scalability as Blockchain networks facing large-scale NLP applications face scalability challenges. Blockchain's intrinsic immutability and transparency raise privacy issues, especially when handling sensitive text data. Private information may be revealed in NLP applications due to the decentralized and publicly accessible nature of blockchain technology. Transparency and trust are fostered by blockchain technology, but there may be trade-offs between the two. Furthermore, not all data can be made public within NLP applications, and it might be difficult to strike a balance between openness and the security of private information.

Our research study reviews the literature, looks at used examples, and analyses obstacles to give readers a thorough grasp of how blockchain technology affects NLP. In the end, this study's unique contribution will be to advance the state of the art in blockchain technology and NLP research, creating chances for cutting-edge applications and encouraging cooperative, safe, and reliable NLP systems.

The integration of NLP with blockchain technologies is examined in the remaining study. The second section gives a general review of NLP and blockchain technology, highlighting both of their key features and range of uses. After that, Section 3 delves into the topic of fusing blockchain technology with natural language processing, examining the possible advantages of doing so. We will examine a range of used cases and applications of blockchain in NLP, such as enhancing text authentication and plagiarism detection, enhancing the origin and validity of data, and providing decentralized translation services. Scalability, privacy, and transparency issues related to this integration will also be looked at, with an emphasis on suggested fixes and future paths. Section four concludes with a list of future projects and conclusions.

27.2 Related Work

The potential for combining blockchain and NLP has been extensively studied in a variety of domains. The authors of [6] proposed a system to improve the security and reliability of event data by taking advantage of blockchain technology's immutability and decentralization features as well as the precision and effectiveness of NLP and RNN algorithms for processing and analyzing natural language data. To guarantee accurate and reliable evaluations, a reputation-based review evaluation system is presented in [7] and is constructed on top of the blockchain system. The reviewers' reputations are then modified depending on this evaluation.

The authors of [8] describe a computational model that uses blockchain technology and natural language processing (NLP) to create intelligent codes from the analysis of legislation. The prototype and preliminary testing of the model are presented, and the results are promising; they demonstrate the applicability of the research topic. Using social media platforms as a key component, the authors of [9] developed a solution for managing disasters and emergency response that focuses on text analysis techniques to enhance the emergency response process of authorities and filter information using automatically gathered data to support relief efforts.

In [10], the research highlights the need for high-quality abstract summaries in handling textual and multimedia content, explaining the difficulties and methods involved. The paper focuses on detecting false media utilizing natural language processing and blockchain technology. The author of [11] presented a system for evaluating employee performance that combines blockchain technology with natural language processing (NLP) to extract characteristics from text input and store them on blockchain, which improves accuracy and transparency. The author of [12] suggested using blockchain technology to securely retain patient records and facilitate quick and easy claims processing. A tabular description of some related research is shown in Table 27.1.

Table 27.1 Various research work on NLP with blockchain.

S. no.	Paper	Authors	Proposed work
1	Blockchain based Secure Event Management System using NLP and RNN Algorithm [6]	Jayakumar D, Haripriya G	This paper, besides using the accuracy and efficiency of NLP and RNN algorithms to process and analyze event data, also uses Blockchain to improve event security by storing validated data in a decentralized, safe manner.
2	Maintaining Review Credibility Using NLP, Reputation, and Blockchain [7]	Zachary Zaccagni; Ram Dantu	Blockchain is used with NLP to preserve review credibility by securely storing reviewers' reputations and evaluation results, thus ensuring reliable evaluations in a decentralized system.
3	Combining Natural Language Processing and Blockchain for Smart Contract Generation in the Accounting and Legal Field [8]	Emiliano Monteiro, Rodrigo Righi	This paper combines NLP and Blockchain for the creation of smart contracts in legal operations. This improves code accuracy and automates the process of interpreting legislation.

(*Continued*)

Table 27.1 Various research work on NLP with blockchain. (*Continued*)

S. no.	Paper	Authors	Proposed work
4	Blockchain-Based Event Detection and Trust Verification Using Natural Language Processing and Machine Learning [9]	Zeinab Shahbazi; Yung-Cheol Byun	In this paper, Blockchain is used in disaster management to verify confidence in event detection through Natural Language Processing. By removing a single point of authority, it improves security and openness and guarantees accurate information sharing on social media.
5	Fake Media Detection Based on Natural Language Processing and Blockchain Approaches [10]	Zeinab Shahbazi; Yung-Cheol Byun	The paper suggests an approach to combine NLP and blockchain with machine learning methods to detect fraudulent accounts and news.
6	Using Natural Language Processing and Blockchain for Employee Performance Evaluation [11]	Umesh Prasad, Soumitro Chakravarty	It combines blockchain and NLP, with the aim to provide a more reliable and accurate way of evaluating employee performance.
7	Smart Patient Records using NLP and Blockchain [12].	Akash Dubey, Kartik Jain	The paper suggests how blockchain can be used to securely keep health records and facilitate quick claim processing.

27.3 Overview on Blockchain Technology and NLP

Due to the potential for influence in various domains, including finance, blockchain technology and natural language processing (NLP) are two emerging technologies. Consequently, investigating the potentially transformational potential of blockchain technology requires an understanding of its foundations and how it intersects with NLP.

27.3.1 Blockchain Technology, Features, and Applications

Blockchain technology, which is a distributed, decentralized ledger that safely records and validates transactions over a network of computers, is frequently linked to cryptocurrencies. It uses cryptographic concepts to guarantee data storage stability, transparency, and trustworthiness. Every transaction or bit of data is included in a "block" that is connected to other blocks to create a chain of information. Due to its numerous distinctive qualities, blockchain technology is becoming more and more popular and widely used in a variety of industries. The features and advantages of blockchain technology are discussed in the following sections.

- **Decentralization**
  One of the fundamental characteristics of the blockchain is decentralization. Blockchain operates on a distributed network of computers, also known as nodes, which collaboratively maintain and verify the integrity of the system and data, in contrast to previous centralized systems [5]. Decentralization also guarantees that no one party has total control over the data, enhancing transparency and lowering the possibility of manipulation. Popular blockchain-based cryptocurrency Bitcoin is operated by a global decentralized network of nodes. Peer-to-peer transactions can be safe and transparent without the need for middlemen when there is no central authority.

- **Transparency and Immutable Records**
  Blockchain keeps an unchangeable, transparent record of all network-stored transactions and data. There is a high level of data integrity and trust since once a transaction or piece of information is recorded in a block, it is nearly hard to change or tamper with. We mention Ethereum (https://ethereum.

org/en/), the second biggest blockchain platform, in this context since it makes smart contract formulation and execution possible. These self-executing agreements are openly documented on the blockchain, guaranteeing the conditions of the agreement's immutability, and fostering mutual confidence amongst the involved parties.

- **Security**
  Blockchain protects data and transactions with cutting-edge encryption technologies. Information is shielded against unwanted access and manipulation by encryption, which also guarantees the information's authenticity, integrity, and confidentiality. To ensure that only authorized parties may access and edit data, a blockchain architecture for business applications, for example, can use encryption techniques to safeguard and validate transactions.

- **Traceability and Audibility**
  Blockchain offers a thorough audit trail of every activity and transaction registered on the network. This traceability feature encourages accountability and transparency by allowing tracking and verification of the ownership, movement, and provenance of assets. In fact, the supply chain sector has improved traceability by utilizing blockchain technology. To trace and validate the provenance and path of food goods, for instance, several food companies employ blockchain technology. This gives customers access to comprehensive details regarding the provenance, ingredients, and processing of a product. The aforementioned features of blockchain technology have the power to transform industries by enhancing efficiency, security, traceability, and transparency. The development of blockchain technology is creating new avenues for innovation and disruption in several industries, including finance, supply chains, healthcare, and government.

27.3.2 Natural Language Processing

The main goal of natural language processing (NLP) is to make computer programs and devices able to communicate with humans through language translation, sentiment analysis, speech recognition, Chabot interactions, and other tasks. For the interpretation, processing, and creation of human language, it combines several technologies, such as rule-based systems, machine learning, deep learning, and language analysis. While these security concerns are rapidly growing, several academics are unaware of the critical role that natural language processing (NLP) approaches play in applications that try to detect spam in mailboxes, intrusion detection, and cyber threats identification. Considering this, we provide a brief description of how NLP approaches are used in spam identification. Spam emails are always changing and no longer fit for old methods of detection. Cybercriminals bypass spam filters by manipulating material and employing social engineering techniques. To detect possible spam, natural language processing (NLP) algorithms examine text patterns, language traits, and email metadata. But to stay put with the latest spams, NLP-based spam detection systems need to upgrade their algorithms frequently.

27.3.3 Challenges in NLP

NLP has numerous difficulties because of the intricacy and ambiguity of human language as well as the rise in data production in the digital age. It is evident that the language used by humans is ambiguous, with words and phrases having several meanings that can be clarified depending on the situation. For NLP systems to accurately comprehend and interpret human language, resolving linguistic ambiguity appears to be a necessary first step. When comprehending a language, knowledge of its syntactic structure and grammatical rules is crucial. It is challenging to capture and represent the intricate syntax and grammar of a natural language because of its variances, exceptions, and informal expressions [19]. Furthermore, the most challenging issue in NLP is determining the semantic meaning of the linguistic components. This implies that a word's meaning might vary depending on the context in which it is used, and that deciphering the intended meaning necessitates a thorough comprehension and study of the context given the precision of that analysis. Large volumes of textual data are needed for NLP systems to train their models and generate predictions. Therefore, maintaining data integrity is crucial to avoiding biases and errors that may result from erroneous, incomplete, or deceptive data gathering. Robust NLP applications require high-quality, properly annotated data sets with data source and validation techniques.

27.3.4 Data Integration and Accuracy in NLP

Because they guarantee dependable outcomes, lessen biases, provide high-quality annotations to data, and create dependable tools, data accuracy and integrity are crucial to NLP. These two elements contribute to the enhanced functionality and efficacy of natural language processing (NLP) systems in various domains, including the healthcare industry, where NLP technologies are essential for processing patient data, including medical reports, health records, etc. In [20] the authors propose an intelligent data integration approach that combines ontology learning, automated mapping, and statistical methods to effectively integrate data and knowledge from diverse sources. Through the use of NLP techniques and semantic analysis, this approach attempts to bridge the gap between disparate data formats and facilitate the integration of valuable information into the knowledge framework.

Since users must have faith in the systems' ability to accurately comprehend and interpret their language input, data accuracy is another essential component in the successful adoption of NLP systems. Moreover, reproducible outcomes, ethical biases or issues that may surface during language processing, and transparent and interpretive natural language processing (NLP) models are all important ways to foster data accuracy. To raise the strength and dependability of NLP datasets, researchers are constantly creating methods for data standardization, cleansing, and annotation. Large numbers of labeled sentences or phrases (related to statistical techniques) are typically included in these datasets, which enable researchers to assess and compare performance.

In NLP, suggested applications and tools must raise the bar for accuracy, data integrity, and cooperation amongst well-established language processing programs. We therefore want to investigate how blockchain and NLP interact. This convergence might offer a viable route for creating innovative solutions that bring together the strength of blockchain technology and natural language processing skills to achieve reliable and safe language processing in the modern digital era.

27.4 Integration of Blockchain into NLP

Combining blockchain technology with natural language processing is a potential field of study and application. NLP researchers can solve numerous issues in this discipline, including data integrity, accuracy, and privacy, by utilizing blockchain technology [21]. Moreover, the decentralized and unchangeable characteristics of Blockchain enable data provenance tracking, guarantee copyright protection, expedite plagiarism detection, and provide safe data exchange among NLP applications.

NLP can validate content originality and safeguard intellectual property (IP) rights by integrating blockchain. Then, because hash values—unique identifiers of the original content—are stored on the blockchain, it is feasible to verify again the authorship and integrity of the textual material. Consequently, this aids in avoiding both plagiarism and infringement of copyright.

Additionally, the blockchain can provide collaborative, decentralized annotation and labeling of textual material, increasing the number of annotated datasets available and enhancing their quality for usage in NLP research projects. Furthermore, Blockchain-based solutions can offer a transparent and safe framework for exchanging NLP data, ensuring that all parties involved may manage their data, giving specific parties access while maintaining data security and privacy. Users can safely exchange resources and language models, for instance, on a blockchain-based data marketplace. One such marketplace would enable businesses and researchers to access and utilize NLP data while retaining ownership and control over their contributions. Furthermore, the construction of reliable linguistic paradigms may be aided by the integration of blockchain technology with natural language processing.

Therefore, users can get increased confidence in the output of NLP models based on the transparency and auditability that blockchain technology provides. Suppose that a blockchain-based language model enables users to track the entire lifecycle of the model, including the used training data and updates this level of transparency ensures that users have visibility into the factors that influence form behavior as shown in Figure 27.2.

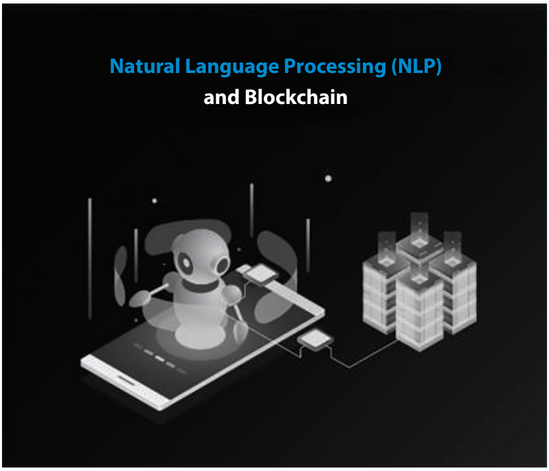


Figure 27.2 NLP and blockchain [13].

27.5 Applications of Blockchain in NLP

There are several opportunities for combining blockchain solutions with NLP. Here are a few examples:

- **Data Source and Authenticity**
  To guarantee that the data used in models and algorithms are dependable and trustworthy, it would be fantastic if Blockchain could be utilized to ascertain the source and credibility of NLP data. We can confirm the accuracy and dependability of the data we work with by logging the source, date of update, and ownership information of the resources on the blockchain. This used case can foster confidence and transparency among the NLP community and aid in the prevention of problems like data tampering, spoofing, and quality concerns. Assume that in order to train AI models, a research team creates a sizable NLP dataset. As a result, they document the steps involved in gathering the data set, such as data sources, pre-processing techniques, and blockchain updates. Blockchain records offer a transparent audit trail that ensures the validity and integrity of datasets shared with other academia or organizations. In this instance, employing blockchain lowers the possibility of using altered or illegal data while enhancing the repeatability and legitimacy of NLP research. It encourages the development of a more reliable and responsible environment for projects involving languages.

- **Distributed Language Resource Sharing**
  Decentralized systems for exchanging linguistic resources, such lexicons and annotated data sets, can be established; thanks to blockchain technology. Through the distributed ledger of a blockchain, NLP researchers can establish a peer-to-peer network that is safe for the exchange of language-related resources. The terms and conditions for resource sharing, such as equity distributions, license agreements, and access rights, can be specified using smart contracts. In this sense, linguists and NLP researchers might establish a blockchain-based platform where they can publish and distribute their multilingual dictionary with other academics

if they collaborate to construct one. Again, in the event that the dictionary is utilized for commercial purposes, smart contracts automatically pay contributors royalties and oversee terms of use and acknowledgement. Collaborating and exchanging knowledge among linguists within the NLP community is facilitated by the sharing of decentralized linguistic resources, which in turn fosters empowerment. In addition, it expedites the creation of more extensive textual materials and lessens reliance on centralized repositories.

- **Enhanced plagiarism detection and text authentication**
  Blockchain technology, which offers an unchangeable record of content that has been timestamped, can improve text authentication and plagiarism detection systems. A document or text becomes nearly impossible to alter or influence covertly when it is stored on a blockchain. This feature can be applied to stop intellectual property theft, identify plagiarism, and confirm the legitimacy of textual content. Taking into consideration a publishing platform that uses blockchain integration to date and store articles that authors submit; in this scenario, each article's hash is recorded on the blockchain, creating an unchangeable record of its existence. By leveraging blockchain for textual resource authentication and plagiarism detection, content providers may firmly ensure the uniqueness and integrity of their written work. As a result, there are more trustworthy publication references, which deters plagiarism and safeguards intellectual property rights.

- **Decentralized Translation Services for Languages**
  Blockchain technology can be used to create decentralized translation services that do not require middlemen and let clients interact directly with qualified translators. Smart contracts can employ user ratings and reputation systems to guarantee quality control [18] if a self-employed translator decides to work as a service provider on a blockchain-based translation platform. Subsequently, they need to make a profile detailing their languages, experience, and previous projects. As a result, clients seeking translation services can check through the translator profiles on the website, evaluate the translators' ratings and testimonials from prior

clients, and select the best translator for their needs. As a result, the platform makes it possible for the client and translator to communicate directly, fostering productive teamwork. Payment settlements will be managed by smart contracts according to predetermined criteria, including word count or project completion benchmarks. The application of Blockchain will have a significant impact on decentralized language translation services, removing the need for middlemen and lowering prices while enhancing process transparency. Additionally, it gives independent translators a stage on which to showcase their abilities and promote themselves to a larger audience. Additionally, user reviews and reputation systems foster trust and assist clients in choosing translators with confidence.

- **Secure and Private Communication:**
  Communication channels that are private and secure can be established with the usage of blockchain technology. Unauthorized access to communications between individuals or organizations can be prevented by encrypting them and storing them on a decentralized network. NLP can be used to preserve communication privacy while analyzing texts and extracting insights like mood or intent.

- **Identity Verification:**
  While natural language processing (NLP) can be used to examine and validate identity information in papers written in natural language, such passports, driver's licenses, and other official documents, blockchain technology can be used to store and authenticate personal identity information. This might contribute to the development of a decentralized, safe identity verification system that is independent of any one government or group.

- **Fraud Detection:**
  While NLP can be used to analyze transactional data and find patterns of fraud or other questionable activity, blockchain development services can be utilized to establish a tamper-proof record of transactions. By doing so, fraud may be avoided, and financial transaction security may be enhanced.

27.6 Blockchain Solutions for NLP

Blockchain technology can be applied to Natural Language Processing (NLP) in many ways. The technologies that can be used to implement blockchain in NLP technologies as shown in Figure 27.2 [13] can be-

Smart contracts: Smart contracts are those that are programmed to run automatically, with lines of code to set parameters of the agreement between the seller and the buyer [16]. They can be used to build decentralized applications that perform functions like voice or text data authentication automatically.

Tokenization: The process of converting data into a digital token that can be traded and transferred on a blockchain is known as tokenization. Tokenization can be used in natural language processing (NLP) to create digital tokens that represent words or phrases that can be kept on a blockchain.

Decentralized storage: Data can be validated and kept on a distributed network by using decentralized storage solutions made possible by blockchain technology. Text or voice recordings, among other types of NLP data, might be safely and impenetrably stored in this way.

Consensus algorithm: Consensus algorithms are employed in blockchain technology to validate transactions and ensure network reliability and security. They could be used in NLP applications to verify the accuracy and legitimacy of data.

Interoperability: The ability of different blockchain networks to exchange data and communicate with one another is referred to as interoperability. For NLP applications that need to transfer data between different networks or systems, this could be helpful as shown in Figure 27.3 and Figure 27.4.

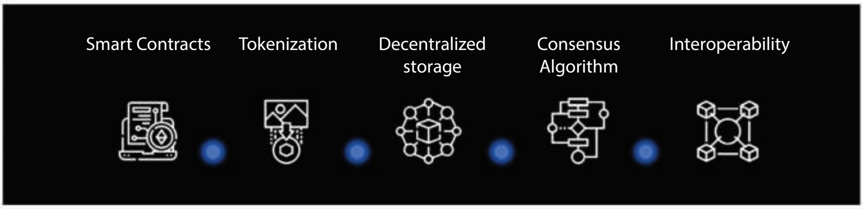


Figure 27.3 Various blockchain solutions.

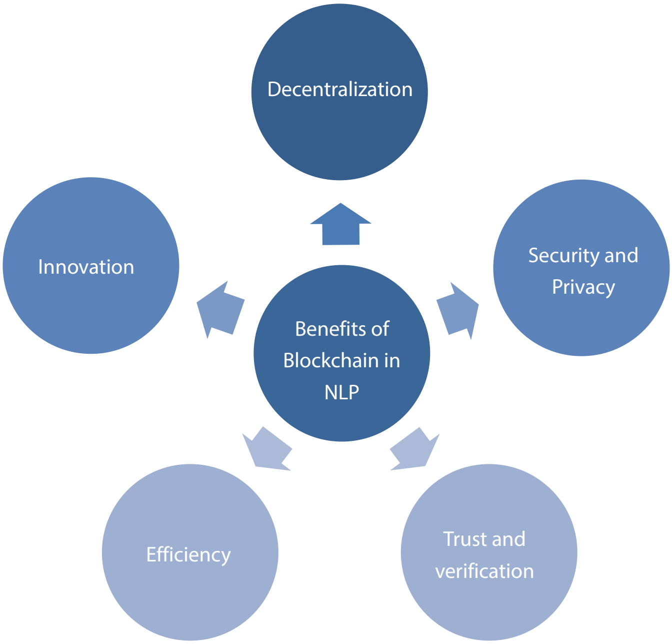


Figure 27.4 Implications of blockchain on NLP.

27.7 Implications of Blockchain Development Solutions in NLP

Blockchain has major implications for NLP and can completely change the way linguistic data is analyzed and handled. The following are some ways that blockchain technology affects NLP:

Decentralization: As blockchain technology lacks a single point of failure, it is inherently decentralized [17]. Because the data is not governed by a centralized authority, NLP systems may be more transparent, safe, and impervious to manipulation.
Security and privacy: To protect language data against compromise, blockchain techniques can be utilized. Sensitive data can be securely and impenetrably maintained on the blockchain, preserving people's privacy and confidentiality.
Trust and verification: Blockchain technology can be used to confirm language data's authenticity and make sure it hasn't been altered. This can help build trust in linguistic data and improve the accuracy and reliability of NLP systems.

Efficiency: Data can be transferred and stored more quickly and securely with blockchain technology, which opens the door to more efficient language processing systems. As a result, processing linguistic data can be completed faster and for less money, with improved accuracy.
Innovation: Blockchain technology and natural language processing (NLP) have the ability to produce novel and creative answers to persistent problems in language processing. By merging the advantages of both technologies, researchers and developers can produce NLP systems that are more precise, effective, safe, and better fit the needs of companies and organizations.

27.8 Sectors That can be Benified from Blockchain and NLP Integration

The integration of blockchain technology with natural language processing has numerous prospects for enhancing a wide array of businesses through the provision of fresh and inventive resolutions to enduring issues [14]. The following are a few sectors that stand to gain from the integration of natural language processing and blockchain technology:

Healthcare: Sensitive patient data can be securely stored using blockchain technology, and insights can be gleaned from the data via analysis and natural language processing. Personalized healthcare solutions that are more successful and efficient could be created using this combination.
Finance: Financial data can be analyzed and interpreted using natural language processing, and financial transactions can be secured by blockchain technology. This combination has the potential to produce financial systems that are more accurate and efficient.
Legal: Legal documents can be securely stored using blockchain technology, and their content can be analyzed, and insights extracted using natural language processing. This combination may be used to produce legal solutions that are more accurate and successful.
Marketing: Natural language processing can be used to evaluate and comprehend client feedback, and blockchain can be used to track and validate the success of marketing campaigns. This could be utilized to develop marketing strategies that are more focused and successful.
Education: Education data may be interpreted using natural language processing (NLP), and educational credentials can be safely stored and verified

using blockchain services. With this combination, educational systems could be developed that are more accurate and productive.

Supply Chain Management: Blockchain is useful for tracking the movement of commodities and ensuring the legitimacy of products while NLP is useful for analyzing and interpreting data from product descriptions, reviews, and user feedback. By combining these two strategies, organizations may establish supply chains that are more responsive to market demands and more transparent and efficient.

Social media and News Analysis: Blockchain technology can be used to confirm the legitimacy of news sources and social media posts While NLP can be used to evaluate and glean insights from the information. This could enhance the information's accuracy and dependability while assisting in the fight against false information and fake news. Thus, businesses and organizations may design more accurate, efficient, and secure systems that better fit their needs by leveraging the characteristics of both technologies.

27.9 Challenges

There can be many challenges to consider when integrating blockchain technology into NLP. Some major challenges are-

- **Scalability and performance:**
  Blockchain technology has scalability issues by nature, particularly public and permission-less blockchains. There is a limit to the number of transactions that can be processed by these blockchains because of the consensus procedures that need nodes to verify and retain every transaction. Scalability becomes an important consideration since NLP [22–25] activities sometimes require processing enormous volumes of data [15]. Complex NLP algorithms can place a strain on the blockchain network [26, 27] due to their increased processing demands, which could result in longer transaction confirmation times and higher transaction fees. Researchers and developers are investigating a range of ways to address scalability difficulties, including layer-two scaling (e.g., off-chain transactions, sidechains), breaking the blockchain up into smaller components, and using more scalable blockchain topologies. By lowering latency and

increasing transaction throughput, these strategies hope to make blockchain-based NLP applications more capable of managing heavier workloads.

- **Implications for privacy and data protection:**
  Privacy and data protection in NLP applications are challenged by the blockchain's transparency and immutability. Transparency can help foster a sense of trust and accountability, but it can also jeopardize the privacy of sensitive information, particularly in situations where privacy is a top priority (such as the healthcare industry). Various techniques, including differential privacy, zero-knowledge proofs, and encryption, can be employed to tackle privacy problems. Privacy can be improved by encrypting data before storing it on the blockchain or by employing zero-knowledge proofs to confirm the legitimacy of data without disclosing its true content. To further address privacy concerns while taking advantage of blockchain technology, privacy-preserving smart contracts can be used, as can private or consortium blockchains.

- **Transparency and Privacy in Blockchain:**
  Integration of NLP Integrating NLP with blockchain presents a significant challenge: striking a balance between privacy and transparency. While every transaction is recorded and verified on the blockchain, which promotes openness, some NLP applications may have privacy concerns that the blockchain cannot address. Techniques like permissioned blockchains and selective disclosure can be utilized to determine a balance. Through selective disclosure, people can share certain information while keeping the rest private. Authorized blockchains give users more control over data exposure and privacy by limiting access to only those who are permitted. The application of blockchain technology to NLP still has a lot of potential to advance data integrity, credibility, and cooperation in this field—despite the obstacles that need to be addressed. Blockchain solutions that are scalable and maintain privacy will be developed through continued study and innovation.

27.10 Conclusion

This study offers a thorough investigation of the combination of NLP with blockchain technologies. This paper advances the fields of blockchain technology and natural language processing by examining use scenarios, evaluating advantages and disadvantages, and outlining potential future research avenues. The findings of this study open new possibilities for creative applications that encourage the use of dependable, safe, and cooperative NLP systems. The future of NLP could be shaped by the combination of both technologies, which could lead to breakthroughs in a variety of industries, such as business, education, and health, as blockchain continues to develop and NLP's capabilities grow.

The field should be advanced, and the issues indicated should be addressed in future work on the blockchain technology and NLP integration. Research might be focused on creating blockchain systems that are scalable and ideal for NLP applications. To ensure that large-scale NLP activities are processed efficiently, this entails investigating alternate techniques and improving transaction throughput. To strike a balance between data privacy and transparency, more research must be done on privacy techniques. The processing of sensitive textual data in blockchain-NLP systems can be done securely and privately because of the developments in privacy-enhancing technologies like differential privacy and secure computing. Additionally, to promote cooperation and integration between various blockchain platforms and NLP systems, efforts should be made to develop interoperability and standards protocols. The promise of blockchain technology to improve natural language processing (NLP) can be further exploited by addressing future research paths, which could lead to revolutionary applications across multiple domains.

References

1. Alphonse Inbaraj, X. and Rama Chaitanya, T., Need to know about combined technologies of Blockchain and machine learning, in: *Handbook of Research on Blockchain Technology*, DOI: https://doi.org/10.1016/B978-0-12-819816-2.00017-4.
2. Salah, K., Rehman, M.H.U., Nizamuddin, N., Al Fuqaha, A., Blockchain for AI: review and open research challenges. *IEEE Access*, 7, 10127–10149, 2019.
3. Integration of AI and Blockchain: All you need to know, https://appinventiv.com/blog/ai-in-blockchain/. (accessed Mar. 18, 2024).

4. Khurana, D., Koli, A., Khatter, K., Singh, S., Natural language processing: State of the art, current trends, and challenges. *Multimed. Tools Appl.*, 82, 3, 3713–3744, 2023.
5. Palmisano, T. and Convertini, V.N., Notarization and Anti-Plagiarism: A New Blockchain Approach. *Appl. Sci.*, 12, 243, 2022, https://doi.org/10.3390/app12010243.
6. Jayakumar, D., Haripriya, G., Ramkumar, M.O., Manjula, S., Blockchain based Secure Event Management System using NLP and RNN Algorithm. *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, 2019, doi: 10.1109/ICAAIC56838.2023.10140877.
7. Zaccagni, Z., Dantu, R., Morozov, K., Maintaining Review Credibility Using NLP, Reputation, and Blockchain. *2022 IEEE 4th International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (TPS-ISA)*, DOI: 10.1109/TPS-ISA56441.2022.00018.
8. Monteiro, E., Righi, R., Kunst, R., da Costa, C., Combining Natural Language Processing and Blockchain for Smart Contract Generation in the Accounting and Legal Field. *2020 International Conference on Intelligent Human Computer Interaction*.
9. Mishra, A.K., Tyagi, A.K., Dananjayan, S., Rajavat, A., Rawat, H., Rawat, A., Revolutionizing Government Operations: The Impact of Artificial Intelligence in Public Administration, in: *Conversational Artificial Intelligence*, pp. 607–634, 2024
10. Nahar, S., Pithawa, D., Bhardwaj, V., Rawat, R., Rawat, A., Pachlasiya, K., Quantum technology for military applications, in: *Quantum Computing in Cybersecurity*, pp. 313–334, 2023
11. Prasad, U., Chakravarty, S., Bisht, Y., Using Natural Language Processing and Blockchain for Employee Performance Evaluation. *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*.
12. Dubey, A., Jain, K., Kalaiselvi, K., Smart Patient Records using NLP and Blockchain. *Proceedings of the 5th International Conference on Smart Systems and Inventive Technology (ICSSIT 2023)*.
13. https://www.lcx.com/natural-language-processing-nlp-and-blockchain/. (accessed Mar. 20, 2024).
14. https://www.oodlestechnologies.com/blogs/blockchain-and-nlp:-uncovering-the-possibilities-and-benefits/.(accessed Mar. 20, 2024).
15. Sikarwar, R., Shakya, H.K., Kumar, A., Rawat, A., Advanced Security Solutions for Conversational AI, in: *Conversational Artificial Intelligence*, pp. 287–301, 2024
16. Song, Z., Shen, P., Liu, C., Liu, C., A Survey on the Integration of Blockchain Smart Contracts and Natural Language Processing. *Proceedings of the 13th International Conference on Computer Engineering and Networks*, 2024.

17. Deshmukh, A., Mishra, A.K., Balamurugan, G., Tyagi, A.K., *Blockchain-Empowered Decentralized Applications Current Trends and Challenges*, chapter 19, Wiley, USA, 2021, https://doi.org/10.1002/9781394213726.ch19.
18. Harris, J.D. and Waggoner, B., Decentralized and collaborative ai on blockchain, in: *Proceedings of the IEEE International Conference on Blockchain (Blockchain)*, 14-17 July 2019, IEEE, Atlanta, GA, USA, pp. 368–375.
19. Pithawa, D., Nahar, S., Bhardwaj, V., Rawat, R., Dronawat, R., Rawat, A., Quantum Computing Technological Design Along with Its Dark Side, in: *Quantum Computing in Cybersecurity*, pp. 295–312, 2023.
20. Kumar, N. and Aggarwal, D., LEARNING-based focused WEB crawler. *IETE J. Res.*, 69, 4, 2037–2045, 2023.
21. Namdev, A., Patni, D., Dhaliwal, B.K., Parihar, S., Telang, S., Rawat, A., Potential Threats and Ethical Risks of Quantum Computing, in: *Quantum Computing in Cybersecurity*, pp. 335–352, 2023.
22. Chirgaiya, S. and Rajavat, A., Tiny object detection model based on competitive multi-layer neural network (TOD-CMLNN). *Intell. Syst. Appl.*, 18, September 2022, 200217, 2023.
23. Rathi, M. and Rajavat, A., *Analysing Cryptographic and Random Data Sanitization Techniques in Privacy Preserving Data Mining*, vol. 83, Allied Publishers, New Delhi, India, 2023.
24. Dhar, S., Dhar, U., Rajavat, A., Factors and Attributes of Team Players: A Study of Engineering Students in India, in: *International Simulation and Gaming Association Conference*, 2021, September, Springer International Publishing, Cham, pp. 53–60.
25. Pithawa, D., Nahar, S., Bhardwaj, V., Rawat, R., Dronawat, R., Rawat, A., Quantum Computing Technological Design Along with Its Dark Side, in: *Quantum Computing in Cybersecurity*, pp. 295–312, 2023.
26. Chauhan, D., Singh, C., Rawat, R., Chouhan, M., Conversational AI Applications in Ed-Tech Industry: An Analysis of Its Impact and Potential in Education, in: *Conversational Artificial Intelligence*, pp. 411–433, 2024.
27. Chauhan, D., Singh, C., Rawat, R., Dhawan, M., Evaluating the Performance of Conversational AI Tools: A Comparative Analysis, in: *Conversational Artificial Intelligence*, pp. 385–409, 2024.

28

Emotion Detection Using Natural Language Processing by Text Classification

Jyoti Jayal[1]*, Vijay Kumar[2], Paramita Sarkar[3] and Sudipta Kumar Dutta[3]

[1]Department of English, SHSS Sharda University, Noida, India
[2]Department of English SOLE, Galgotias University, Noida, India
[3]Department of C.S.E JIS UNIVERSITY, Nilgunj Road, Agarpara, Kolkata, India

Abstract

Emotion recognition is one of the branches within the field of emotion analysis. Advanced emotion analysis is required to identify a sentence as positive, negative, or neutral. Our emotions get more profound as we comprehend the positive or negative tone of the statements or ideas. People communicate their emotions using various methods, such as facial expressions, written language, verbal communication, and bodily movements. Opinions and perspectives form the foundation of individuals' emotions, actions, and the impact of those emotions on their speech and actions. Linguistics and machine learning are broad areas that include subfields like computational linguistics, natural language processing (NLP), and artificial intelligence (AI). Thanks to Natural Language Processing (NLP), computers are now capable of comprehending, analyzing, and interpreting spoken language. Natural language processing (NLP) has become an essential component of our daily lives due to the widespread adoption of machine translation programs, voice assistants, and search engines. This article examines various machine learning and deep learning techniques to the task of emotion recognition through text classification utilizing natural language processing.

***Keywords*:** Emotion detection, natural language processing, machine learning, deep learning, ensemble learning, text classification

**Corresponding author*: jyoti.jayal@sharda.ac.in

Rajesh Kumar Chakrawarti, Ranjana Sikarwar, Sanjaya Kumar Sarangi, Samson Arun Raj Albert Raj, Shweta Gupta, Krishnan Sakthidasan Sankaran and Romil Rawat (eds.) Natural Language Processing for Software Engineering, (425–442) © 2025 Scrivener Publishing LLC

28.1 Introduction

Emotional expression is a fundamental and intrinsic component of human behavior. Essentially, they are embodiments of the distinct feelings and experiences influenced by both internal and external causes. Emotions can be represented through several means, such as facial expressions, nonverbal indications during conversations, body posture and language, physiological responses, and written words. The importance of text as a means of communication has significantly increased as a result of the widespread use of social media applications and platforms. In contemporary times, the majority of individuals engage in communication through various digital platforms such as text messaging, tweets, online posts, blogs, and so on. Consequently, we now have access to a vast amount of social data, which presents numerous intriguing opportunities for a variety of applications [1].

Analyzing the emotional state of content shared on social media sites has various potential applications, including the development of affect-aware human-computer interfaces. Additional possible applications involve monitoring corporate reputation and brand management. Psychologists could employ an emotion detection system to enhance their comprehension of how to assist individuals when they articulate their concern orally, or to ascertain the intended tone of an email prior to its transmission. By employing emotion recognition on textual data extracted from popular social media platforms, candidates running for public office can effectively customize their political agenda to align with the desires and preferences of the voting population.

Precise and methodical emotion recognition aids in closing the communication divisibility between computers, which have difficulty understanding emotions, and individuals who are very expressive. The primary objective of affective computing is to develop computational systems that adapt their response according on the user's emotional state. For a system interface to effectively respond to users' different emotional states, it is essential for it to reliably recognize their affective or emotional states. Therefore, the detection of these states is of utmost importance. Enhancing the computer-user interaction can be achieved by automatic emotion recognition, which enables an interface to respond to the emotional states of the user. Interfaces that are sensitive to users' emotions can considerably enhance both human-computer interaction (HCI) and computer-mediated communication (CMC). Research into emotion analysis is necessary for the development of text-to-speech (TTS) synthesis systems [2].

Emotion-aware text-to-speech systems possess the capability to detect nuances in textual content and enhance the translation to get a more authentic and genuine sound. Various disciplines are currently engaged in the development of affect-sensitive interfaces, including educational technology, mental health, and gaming. Developers can utilize emotion recognition to customize their systems according to the user's emotional state and establish automated responses based on this data. Recent advancements in AI and ML have opened up new opportunities to include emotional sensitivity and expressiveness into HCI interfaces.

Emotions are impacted by multiple factors and have their roots in cognition. Individuals convey their emotions by means of their facial expressions and spoken or written communication. Our approach solely focuses on recognizing the emotional orientation of textual utterances made on different social networking platforms. The subjective human appraisal determines the emotional content and the character of the exhibited emotion. It is widely recognized that, just as there are multiple ways to convey emotions, different readers may interpret the same expression in various ways. The approaches given here were tested using data that had an emotional orientation and was rated as identical by at least two persons. The main objective of affect detection research has been to determine the positive, negative, or neutral polarity of sentiments. This study specifically examines one aspect of affect identification that has received less attention: the ability to recognize emotions that are relevant to a certain domain. Through acquiring the skill to discern the emotions conveyed in written material, readers can enhance their understanding of the author's objectives, attitude, and the content of the text.

28.2 Natural Language Processing

Computational linguistics, natural language processing (NLP), and artificial intelligence (AI) are subdisciplines that fall within the broader sciences of machine learning (ML) and linguistics. Natural Language Processing (NLP) enables computers to understand, analyze, and interpret spoken language. Natural language processing (NLP) has become an essential component of our daily lives due to the widespread use of machine translation programs, voice assistants, and search engines. Identifying the semantic context of words and managing their diverse interpretations within different sentences in a document or corpus are just two of the numerous challenges in natural language processing. Some of the challenging tasks include: recognizing sardonic or caustic phrases; identifying words with contradictory information; and understanding statements with several interpretations. Different NLP approaches are presented in Figure 28.1.

Rule-based Approaches - Patterns are frequently analyzed or compared using rule-based methods. Generally speaking, these approaches can be quite effective for certain jobs. However, they deteriorate when employed on a worldwide level.

Traditional machine learning Approaches - Conventional machine learning relies on training models and deriving conclusions from test sets as its primary methods. By employing this approach, it is possible to achieve many natural language processing tasks, such as sequence tagging. Models are utilized to derive meaningful inferences from the data.

Neural network Approaches - Word embeddings, which are word vector representations, serve as inputs for training neural networks. Various alternative neural network architectures, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks, have been extensively researched. Additionally, there have been encouraging advancements in integrating deep learning with natural language processing.

The significance of unstructured data in processing, analysis, and decision-making is being demonstrated by the rapid increase in their abundance. When managing vast quantities of significant content such as biographies, photographs, videos, and tweets, natural language processing (NLP) is the optimal choice. Advanced AI systems often incorporate Natural Language Processing (NLP) and Machine Learning (ML) methods to enhance their ability to perform tasks [3]. Siri and Alexa, two AI-based virtual assistants, utilize natural language processing (NLP) techniques and

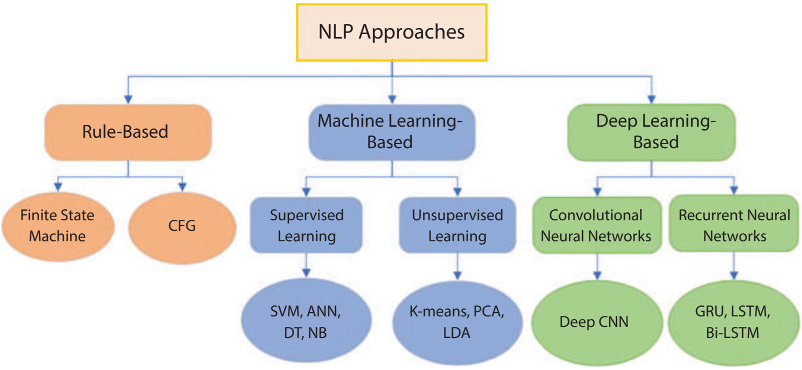


Figure 28.1 Different NLP approaches.

voice recognition to comprehend human speech. They leverage machine learning algorithms to analyze patterns learned from previous interactions.

Some of the applications of NLP are listed below:

- Language Translation
- Social Media Monitoring
- Chat bots and Virtual assistants
- Survey Analysis
- Voice Assistants
- Grammar check
- E-mail Filtering
- Market Intelligence
- Speech Recognition
- Text Extraction
- Text Classification
- Text Summarization
- Web Sampling
- Face Recognition
- Pattern Recognition

28.3 Emotion Recognition

Emotion analysis includes various subfields, and one of them is emotion recognition [6]. Advanced emotion analysis can only go as far as classifying a statement as positive, negative, or indifferent. Once we comprehend the negative or good nature of the statements or thoughts, our emotions delve into a more profound level. Facial expressions, written words, verbal communication, and physical gestures are all means through which individuals express their emotions. sentiments, behaviors, and the consequences of those sentiments on people's words and actions are all based on personal opinions and perspectives [4].

The main purpose of this tool is to assist businesses in assessing the sentiment of their customers towards their brand or product. Within the medical domain, it possesses the capacity to mitigate patients' requirements for medication. One possible use in the field of automation is the ability to understand human emotions. When a driver experiences drowsiness, an emotion recognition system can alert them. Emotion analysis and recognition is a broad field of research that offers useful understanding of spoken and written language. Plutchik's proposed approach of mood swings allows for the differentiation of several emotions, including rage, fear, melancholy, disgust, surprise, desire, optimism, and joy.

Investigating human emotions poses significant challenges. Many corporations and groups utilize it due of its immense power. The surge in popularity of micro-blogs in recent years has led to the emergence of a new area of research that focuses on automatically recognizing emotions in these entries. Micro-blog postings, due to their brevity and lack of structure, cannot be readily processed like longer texts often used in text mining studies. Microblogs use several symbols, conventions, and unique usages that can greatly influence the impact of the text. Psychologists and behavioral scientists have dedicated significant time and effort to studying emotions due to their fundamental importance in the human experience. Computer scientists specializing in human-computer interaction have also shown a keen interest in emotion analysis. They conduct studies on facial expressions, speech recognition, and emotion recognition using various methodologies [5].

28.4 Related Work

28.4.1 Emotion Detection Using Machine Learning

A novel proposition has been put out to categorize emotions with machine learning methodologies that rely on physiological attributes. Jang *et al.* [6] employed four machine learning methods, namely LDA, CART, SOM, and SVM, to categorize the emotions of 200 college students, specifically boredom, pain, and surprise. Physiological signals, such as electrodermal activity (EDA), electrocardiogram (ECG), photoplethysmography (PPG), and skin temperature (SKT), were measured for one minute before the emotional test. The differential values of each attribute were subtracted from the baseline to categorize emotions. Emotions served as the fundamental basis for machine learning algorithms. Based on the findings, SVM classification yielded the most optimal results, and emotional accuracy is of utmost importance. Many human emotions can be inferred from physiological indicators.

Psychologists have been intrigued by research in neuroscience, linguistics, and human-computer interaction systems that focus on autonomous face recognition and facial emotion recognition.

Jaikrishnan [7] developed a recognition system called AFERS, which consists of three steps: face recognition, feature extraction, and detection of facial expressions. The YCbCr color scheme was used to ascertain an individual's skin tone. Lighting compensation was used to establish consistency in the face, while morphological operations were employed to preserve the desired appearance. The data extraction process utilized the outcome of the initial step. The Active Augmentation Method (AAM) was used to augment the eyes, nose, and mouth. The Appearance Model was utilized to calculate an individual's physical appearance. The final stage involved automated facial recognition.

The intensity of textual emotions can be determined by assessing them, as stated by Mohammad and Bravo-Marquez [8]. The tweets, when analyzed with annotations, displayed intense emotions such as anger, fear, joy, and sadness. The writers enhanced annotation consistency and provided reliable fine-grained scores by implementing best-worst scaling (BWS).

Tsai and Chang [9] developed a three-stage support vector machine (SVM) to detect facial expressions. Each of the twenty-one support vector machines (SVMs) used in the initial phase consisted of a binary amalgamation of seven phrases. The initial phase would suffice if there were just two phrases; the subsequent step would be utilized if there were three; and the final stage would be utilized if there were all three.

Healy *et al.* [10] has published a full account of their novel method for identifying emotions in real-time video. The work showcased the implementation of a distinct machine learning support vector machine (SVM) for rapid and precise classification, utilizing 68-point facial features. Through a regulated laboratory environment, the application has acquired the ability to recognize six distinct emotional states. The tracking of changes in facial expressions was beneficial in assessing individuals' emotional states in the given settings.

Tsai and Chang [9] proposed a FERS approach that utilized SQI filters to extract Haar-like characteristics from images. There was an enhancement in the rate of face detection. SQI filters are renowned for their exceptional effectiveness in various lighting conditions. The discrete cosine transform (2D-DCT) and Gabor filtering have been chosen to represent the features. Finally, support vector machines (SVMs) were used as the classifiers to assign various categories to the facial expressions. Schneider *et al.* [11] utilized the Wall Street Journal (WSJ) test set to showcase a novel application of wav2vec in unsupervised pre-training for speech recognition. Their approach yielded an impressive Word Error Rate of 2.43%. Baevski and colleagues [12] developed Vq-wav2vec as a means to examine discrete representations of audio samples using the Timit corpus as a benchmark. This approach utilizes a self-supervised prediction task that leverages BERT pre-training archives.

Deshmukh *et al.* [13] used feature extraction and classifier training to recognize emotions from audio data. The feature vector encompasses the elements that characterize the audio signal. When evaluating a speaker, it is important to consider their intensity, pitch, and tone. Hence, it is crucial to instruct the classifier model in recognizing the emotional expression of a certain object. The process of segregating the source dataset from the training and testing datasets was performed manually. Anger, despair, and joy were recovered as Mel-Frequency Cepstral Coefficients (MFCC) from the audio.

Deng and Ren [14] proposed a Multi-Label Emotion Detection Architecture (MEDA) to accurately identify all the emotions expressed in a given text. The fundamental components of MEDA consist of the Emotion Correlation Learner (ECorL) and the MultiChannel Emotion-Specific Feature Extractor (MC-ESFE). By utilizing its MC-ESFE module, which consists of several ESFE networks for each channel, the system successfully extracted certain features related to underlying emotions.

Uddin and Nilsson [15] proposed a dependable method for emotion recognition by utilizing audio speech as a dataset for machine learning. The computation of Mel-frequency cepstral coefficients (MFCC) was used as a feature to train emotion recognition algorithms that are not dependent on individuals. Audio data is necessary for this purpose. It is crucial to acknowledge and distinguish the following emotions: sadness, wrath, scorn, contempt, surprise, and terror. [16] created a surveillance system to identify these emotions in video recordings. The target demographic of the product was the elderly.

Tafreshi and colleagues [17] conducted research that investigated the impact of news-based essays on collaborative tasks related to empathy and emotion prediction. The combination of features extracted from transformer models with concurrent learning problems allows for strong modelling of empathy, discomfort, and emotion inside a transformer language model. Despite extracting lexical features from sentiment, opinion, and emotion lexicons, these resilient contextualized aspects have a significant impact on empathy, distress, and emotion models.

28.4.2 Emotion Detection Using Deep Learning

The advancement of efficient classifiers for speech emotion recognition has predominantly depended on models that utilize audio data. To enhance comprehension of voice data, Yoon and colleagues [18] developed a novel deep dual recurrent encoder model that concurrently utilized both textual data and audio inputs. This model employed dual recurrent neural networks (RNNs) to capture information from both audio and text sequences to predict the emotion category, taking into account the auditory and linguistic aspects of emotional speech. This architecture differs from models that solely focus on auditory elements by thoroughly analyzing speech data at all levels, encompassing signals and languages. As a result, it effectively utilizes the intrinsic information included in the data. The efficacy and attributes of the proposed model were evaluated through extensive experiments. When tested on the IEMOCAP dataset, this model demonstrated accuracies ranging from 68.8% to 71.8% in accurately categorizing data into four emotion categories: angry, joyful, sad, and neutral. This outperformed prior cutting-edge methods.

Yang *et al.* [19] introduced a face recognition-based method to assess a student's understanding of distant learning. The researchers proposed a three-step methodology for constructing emotion identification models: feature extraction, subset feature selection, and emotion classification. The input image depicted a human face, which was identified using a Haar Cascades methodology. The eyes and lips were isolated using Sobel edge detection. Subsequently, the distinctive value was acquired. The training of a Neural Network classifier resulted in the identification of six unique emotional groups. The JAFF database experiments confirmed the method's exceptional classification performance. Experimental results indicate that students' expressions in virtual learning settings align with this approach. This study demonstrates the practicality of employing facial expression-based emotion identification to identify a student's real-time learning state in remote education. Hence, instructors should modify their instructional approaches in online classrooms according to the emotional states of their pupils, which can be beneficial.

Ortis, Farinalla, and Battiato [20] offer a thorough examination of the current body of research on image sentiment analysis by discussing various methodologies, datasets, and challenges that new researchers may face. The article states that emotional models, Image Sentiment Analysis datasets, and feature design are the three fundamental elements for constructing ISA systems. They provided comprehensive explanations for each of these choices, utilizing cutting-edge research to support their assertions. Social networking platforms, which users frequently utilize to upload and share photographs, offer convenient access to datasets like VSO and others. Recent discoveries in picture sentiment analysis indicate that models for this particular type of analysis experience advantages when employing learning approaches such as multi-modal embedding and convolutional neural networks. The discussed topics encompassed picture popularity and virality prediction, relative attributes, sentiment and ideograms, as well as the challenges faced by academics.

[21] developed a multimodal emotion recognition model that utilizes both text and speech as inputs. Haryadi and Kusuma [22] employed deep learning techniques, specifically Long Short-Term Memory (LSTM) and Nested Long Short-Term Memory, to classify seven emotions (angry, fear, joy, love, sadness, surprise, and thankfulness) in text. They trained a Bi-LSTM (bidirectional long short-term memory) network that utilizes captured textual features as fusion features. Utilizing the Twitter API, a dataset was generated, subjected to pre-processing, and subsequently evaluated by comparing the results with Support Vector Machines (SVM). SVM utilized TF-IDF as features for classification. After training the LSTM and Nested LSTM models, we collected the accuracy and loss outputs from the checkpoints. Nested LSTM demonstrates superior performance compared to the other two approaches, with an accuracy rate of 99.167 percent.

[23] presented a classification of expressions based on the emotions of joy, fury, fear, and melancholy. A deep Convolutional Neural Network (CNN) has been employed, utilizing learned word vectors, to categorize the provided dataset. This approach suggests four unique categories, namely joy, anger, fear, and sadness, to streamline the process of classifying emotions in Arabic tweets. Their proposed approach for sentence classification problems employed deep Convolutional Neural Networks (CNNs) that were trained using dataset-specific word vectors. To evaluate the EI-oc aim, they assessed the effectiveness of their proposed deep learning approach by conducting experiments on the SemEval Arabic tweets dataset. The primary procedures involved in this task were the vectorization of words, phrases, documents, and classification. This study investigated the performance of three different models: CNN-static, CNN-non-static, and CNN-rand. The outcome was compared using three additional machine learning algorithms: Support Vector Machines (SVM), Naive Bayes (NB), and Multilayer Perceptron (MLP). Three different Arabic stemmers, namely Light stemmer, ISRI, and Snowball, were used to implement these methodologies. In addition, two primary feature variables were utilized: Count and TF-IDF.

Lopez and Ivan [24] conducted a study in the United States to investigate public sentiment towards the COVID-19 vaccination. A total of 2,000 tweets were extracted from Twitter for the dataset between February 3, 2021, and February 10, 2021. The data was categorized based on factors such as location, description, sentiment, number of followers, number of likes, and other relevant criteria. A preliminary analysis of the collected tweets using vaccine-related keywords was performed and a sentiment score ranging from -2 to +2 was employed to reflect the sentiment. A statement was deemed positive if its value above zero, and negative if its value fell below zero. Before evaluating tweets based on their creation date and location, the dataset was organized based on the tweets with the highest number of favorites. There was a mixture of positive and negative comments about vaccination, with somewhat more positive tweets than negative ones.

28.4.3 Emotion Detection Using Ensemble Learning

Perikos and Hatzilygeroudis [25] implemented a classifier ensemble strategy utilizing bagging and boosting methods to successfully identify emotional material from social media. The experiments made use of two datasets: the International Survey on Emotion Antecedents and Reactions (ISEAR) and the Affective text dataset. The two datasets underwent analysis using machine learning classifiers known as Naive Bayes and Maximum Entropy. The results demonstrated that the proposed classifier significantly improved Twitter's ability to recognize emotions.

Po-Yuan Shih and Chia-Ping Chen [26] proposed a model that utilizes ensemble learning techniques with neural networks to enhance the accuracy of voice emotion identification tests. This model employs an imbalanced dataset to train subsets that are balanced, and subsequently merges them to generate predictions. By utilizing the FAU-Aibo database, experiments were conducted and determined that a recall rate of 45.5% produced the most optimal outcomes. This yielded innovative results in comparison to the old model. Rather than using a large, unbalanced dataset, this result was achieved by using smaller, more balanced subsets of the data.

Aribisala and colleagues [27] developed a model for emotion recognition using EEG data. To ascertain emotions, our model utilized a publicly available EEG dataset. Data from participants were gathered by assessing four emotional states: arousal, dominance, valence, and preference. The dataset yielded three features: energy, wavelet entropy, and standard deviation. Regarding sensitivity, specificity, and accuracy, this model demonstrated superior performance with scores of 97.54%, 99.21%, and 97.80%, respectively.

[28] suggested a method for face expression regression employing multilevel features in a Convolutional network. The researchers attributed certain emotions such as surprise, neutrality, sadness, fear, anger, happiness, and disgust to distinct face regions in the photos. The FER2013 dataset was used to conduct the experiments. In comparison to the most advanced methods, their model demonstrated promising outcomes.

Mishra *et al.* [29] proposed a method for classifying emotions based on EEG patterns. Their approach involved combining a support vector machine with an ensemble Convolutional network. The process entailed utilizing deep learning models to extract features, followed by employing a support vector machine classifier to accurately categorize EEG patterns into their respective emotion classifications. The emotional state of the subject was determined by employing a Support Vector Machine classifier that considered both single and multiple emotional characteristics. In this experiment, physiological signals (DEAP) were utilized to assess emotions using the most renowned convolutional networks, namely AlexNet and GoogleNet. while categorizing a single attribute (valence), the accuracy achieved using AlexNet and SVM was 87.5%. However, while classifying two attributes (arousal and valence), the accuracy rate dropped to 62.5%.

Youngquist [30] developed a novel neural network architecture for classifying emotions based on short sentences. In this approach, a combination of recurrent, convolutional, and pooling layers was employed to effectively capture the emotional context of the text. Emotions were extracted from the text by conducting experiments utilizing five separate data sets. Compared to the historical models and all data sets, this technique produced an average enhancement rate of 8.31% based on the experimental findings. This ensemble recurrent convolutional network utilized only few sentences as the sole input for emotion classification.

Salama *et al.* [31] proposed the design of a 3D-Conventional Neural Network to extract spatio-temporal properties and video data from human faces. The final fusion productions were decided by integrating data augmentation with ensemble learning methodologies. The system employed RCNN object detection techniques in combination with the OpenCV libraries to precisely extract the facial characteristics related to emotional content. Two fusion approaches, bagging and stacking, were employed in the conducted experiments. The stacking technique resulted in the highest recognition accuracy rate of 96.3 percent and valence accuracy of 96.7 percent. The utilization of the grid search ensemble learning technique facilitated the achievement of this degree of precision.

(Sun *et al.*) Modified iterations of gradient-based explanation techniques and Layer-wise Relevance Propagation (LRP) were employed to incorporate attention mechanisms into photo captioning models. The focus of the study was to analyze the predictions made by these models, surpassing mere attention visualization. This study thoroughly compared the interpretability of attention heat maps using methods such as LRP, Grad-CAM, and Guided Gradient-weighted Class Activation Mapping (G-GAM).

Robots can derive advantages from Wisha Zehra *et al.*'s [32] ensemble learning-based cross-corpus multilingual speech recognition system, which enables them to provide more efficient and emotionally expressive responses. This model compared ensemble learning with more traditional machine learning methods. Through experimental research, it was discovered that several classifiers achieved the highest level of accuracy across different corpora. Ensemble learning was employed to amalgamate many classifiers. The trials revealed a 13% enhancement in accuracy for the Urdu corpus, an 8% enhancement for the German corpus, an 11% enhancement for the Italian corpus, and a 5% enhancement for the English corpus. The experimental results repeatedly showed that the innovative ensemble learning strategy considerably enhanced accuracy in comparison to state-of-the-art methods.

28.5 Machine Learning Techniques for Emotion Detection

Regarding data mining, the KNN algorithm is the most straightforward approach for managing classification tasks and regression problems. It utilizes a non-parametric approach to arrange newly acquired data by comparing it to similar data in the training set. The KNN approach, being a collaborative learning strategy, bypasses the process of constructing models prior to testing them. Understanding the correlation between the attributes and the outcomes is not a prerequisite for the effectiveness of this technique; it is highly flexible. During a traffic gridlock, the nearest neighbor value can replicate the current condition by recreating the corresponding historical data. A KNN-based forecasting approach can be constructed using three primary components: a monitoring database, a nearest-neighbor finding procedure, and a forecasting task. The conventional approach to identifying similarities across records involves calculating the edit distance (ED). The KNN method can be employed to estimate the speed and flow of data based on the inputs of density, with density being the desired result. This approach is used to identify nearby and essential input data. This function retrieves the closest objects by calculating their Euclidean distance whenever a data point is given. It is possible to calculate, provided that there is a sufficient amount of data available.

The CNN technique primarily focuses on managing 2D data, such as photographs. The model consists of several concealed layers, such as convolution layers, fully linked layers, and pooling, along with an input and output layer. A convolutional layer is capable of retaining features by applying filters that modify the input value according to a predetermined pattern. The next phase involves utilizing a minimum or maximum matrix to combine the clustering outcomes from the previous stage into distinct sections at the subsequent level. The pooling layer simultaneously decreases dimensionality to assist the overall model and imparts knowledge of abstract level data.

The Artificial Neural Network (ANN) is widely recognized as the prominent model for predicting traffic flow. Some of its strengths include multidimensional value management, adaptability, learning, and increased generalization. This artificial neural network (ANN) technique excels at dealing with input values that are noisy or absent, and it eliminates the need for data estimation. Typically, it is defined by employing three types of identifiers and constructed as a multilayer network system. The parameters encompass the arrangement of communication between the layers, the mechanism for initiating input-output, and the frequency at which the layers' weights are updated. Artificial Neural Networks (ANN) are considered the fundamental basis of Artificial Intelligence (AI). They are a non-parametric approach to solving problems. Neural networks (NN) are trained to execute a pre-established task by regulating the weights of the connections between the various units. These units operate in coordination. "Neurons" refer to the individual components responsible for processing information. The regulation of neuron relevance is necessary for training the neural network. Each individual neuron utilizes a link to communicate a scalar value p for intake. We have to compute the product of the input value and its weight, add the bias data, and subsequently use a transfer mechanism to derive the target value. The function f takes this value as input and produces the result as output. To achieve the desired result, adjustments are made to the bias value (b) and the weights (w). The neuron's output is the input value that is considered by other neurons, and this process is iterated for each data point. Sigmoid, purelin, hardlim, and other transmission mechanisms are available options for function f, depending on the specific condition needed.

The multilayer perceptron is a type of artificial neural network that utilizes feed-forward connections. The primary strengths of this model lie in its numerous layers and its capacity to effectively process non-linear input. The input layer, consisting of seven nodes, receives data from the user and transmits it to the maximum likelihood prediction (MLP) algorithm. This is achieved by performing the dot function on each layer and combining it with the weight value of the layer below. Except for the output type layer, the activation method is employed to transmit output using the dot function in all other layers. Once the output type layer contains only one node, the process is iterated for all layers. The anticipated velocity corresponds to the value obtained from a single node. Every layer in this paradigm relies on and is influenced by every layer that is positioned beneath it. During the training process, the error value is calculated by comparing the real value with the output values. The erroneous data is inputted into the Rprop method, which specifically aims to reduce the magnitude of the partial derivative. The magnitude of the weight is influenced by its new value, whereas the sign of the derivative function indicates the specific direction of the weight.

28.6 Conclusion

To attain a more genuine and realistic tone, text-to-speech algorithms that are capable of detecting emotions should be able to perceive nuances in the text and enhance the output accordingly. Currently, various industries such as educational technology, mental health, and gaming are actively developing interfaces that are capable of detecting and responding to human emotions. Emotion detection enables programmers to customize their systems based on the individual user's distinct emotional state and establish automated responses accordingly. Recent advancements in machine learning (ML) and artificial intelligence (AI) have opened up new opportunities to integrate emotional sensitivity and expressiveness into human-computer interaction (HCI) interfaces. This article discusses various machine learning and deep learning strategies for emotion recognition via text classification using natural language processing.

References

1. Kawade, O. and Oza, K., Sentiment Analysis: Machine Leaming Approach. *Int. J. Eng. Technol.*, 9, 2183–2186, 2017, I 0.21817/ijel/2017/v9i3/l 709030151.
2. Mishra, A.K., Tyagi, A.K., Dananjayan, S., Rajavat, A., Rawat, H., Rawat, A., Revolutionizing Government Operations: The Impact of Artificial Intelligence in Public Administration, in: *Conversational Artificial Intelligence*, pp. 607–634 , 2024.
3. Drus, Z. and Khalid, H., Sentiment Analysis in Social Media and Its Application: Systematic Literature Review. *Procedia Comput. Sci.*, 161, 707–714, 2019, 10.1016/j.procs.2019.11.174.

4. Nahar, S., Pithawa, D., Bhardwaj, V., Rawat, R., Rawat, A., Pachlasiya, K., Quantum technology for military applications, in: *Quantum Computing in Cybersecurity*, pp. 313–334, 2023.
5. Raghuvanshi, A., Veluri, R. *et al.*, Learning analytics using deep learning techniques for efficiently managing educational institutes. *Mater. Today: Proc.*, 51, 2317–2320, 2022, Available: 10.1016/j.matpr.2021.11.416.
6. Jang, E., Park, B., Kim, S., Sohn, J., Emotion classification by machine learning algorithm using physiological signals. *Proc. Comput. Sci. Inf.*, vol. 25, pp. 1–5, 2012, [Online].Available http://www.icmlc.org/icmlc2012/00I_icmlc2012.pdf.
7. Jaikrishnan, V., Emotion Detection Using Machine Learning. *Int. J. Recent Trends Eng. Res.*, 3, 6, 28–32, 2017, https://doi .org/10.23883/ijrter.2017.3267.lovl p.
8. Mohammad, S. and Bravo-Marquez, F., Emotion Intensities in Tweets, pp. 65–77, 2017, 10.18653/v1/$17-1007.
9. Tsai, H.H. and Chang, Y.C., Facial expression recognition using a combination of multiple facial features and support vector machine. *Soft Comput.*, 22, 13, 4389–4405, 2018, doi: 10.1007/s00500-017-2634-3.
10. Healy, M., Donovan, R., Walsh, P., Zheng, H., A Machine Leaming Emotion Detection Platform to Support Affective Well Being. *Proc. - 2018 IEEE Int. Conf. Bioinforma. Biomed. B1BM 2018*, vol. I, pp. 2694–2700, 2019, doi: IO.1109/B1BM.2018.8621562.
11. Schneider, S., Baevski, A., Collobert, R., Auli, M., WAV2vee: Unsupervised pre-training for speech recognition. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019-Septe, pp. 3465–3469, 2019, https://doi.org/10.21437/ Interspeech.2019-1873.
12. Baevski, A., Schneider, S., Auli, M., vq-wav2vec: Self-Supervised Learning of Discrete Speech Representations, pp. 1–12, 2019, http://arxiv.org/abs/1910.05453.
13. Deshmukh, G., Gaonkar, A., Golwalkar, G., Kulkarni, S., Speech based emotion recognition using machine learning. *Proceedings of the 3rd International Conference on Computing Methodologies and Communication, ICCMC 20/9*, Jccmc, pp. 8 I 2–8 I 7, 2019, https://doi.org'I 0.1I09/ICCMC.2019.88 I 9858.
14. Deng, J. and Ren, F., Multi-label Emotion Detection via EmotionSpecified Feature Extraction and Emotion Correlation Leaming. *IEEE Trans. Affective Comput.*, 3045, 0, 1–12, 2020, https://doi.org/10.1109/TAFFC.2020.3034215.
15. Uddin, M.Z. and Nilsson, E.G., Emotion recognition using speech and neural structured learning to facilitate edge intelligence. *Eng. Appl. Artif. Intell.*, 94, September, 103775, 2020, https://doi.org/10.1016/j.engappai.2020.103775.

16. Shrivastava, M., Patil, R., Bhardwaj, V., Rawat, R., Telang, S., Rawat, A., Quantum Computing and Security Aspects of Attention-Based Visual Question Answering with Long Short-Term Memory, in: *Quantum Computing in Cybersecurity*, pp. 395–412, 2023.
17. Tafreshi, S., De Clereq, O., Barriere, V., Buechel, S., Sedoc, J., Balahur, A., W ASSA 2021 Shared Task: Predicting Empathy and Emotion in Reaction to News Stories. *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 92–104, 2021, https://aclanthology.org/2021. wassa-1.10.
18. Yoon, S., Byun, S., Jung, K., Multimodal Speech Emotion Recognition Using Audio and Text, 2018, I 0.1109/SLT.2018.8639583.
19. Yang, D., Alsadoon, A., Prasad, P.W.C., Singh, A.K., Elchouemi, A., An Emotion Recognition Model Based on Facial Recognition in Virtual Leaming Environment. *Procedia Comput. Sci.*, 125, 2009, 2–10, 2018, https://doi.org/10.1016/j.procs.2017.12.003.
20. Ortis, A., Farinella, G.M., Battiato, S., An overview on image sentiment analysis: Methods, datasets and current challenges. *ICETE 2019 Proceedings of the I6th International Joint Conference on e-Business and Telecommunications*, 2019.
21. Sikarwar, R., Shakya, H. K., Kumar, A., Rawat, A., Advanced Security Solutions for Conversational AI, in: *Conversational Artificial Intelligence*, pp. 287–301, 2024.
22. Haryadi, D. and Kusuma, G.P., Emotion detection in text using nested Long Short-Term Memory. *Int. J. Adv. Comput. Sci. Appl.*, 10, 6, 351–357, 2019, https://doi.org/10.14569/ijacsa.2019.0100645.
23. Pithawa, D., Nahar, S., Bhardwaj, V., Rawat, R., Dronawat, R., Rawat, A., Quantum Computing Technological Design Along with Its Dark Side, in: *Quantum Computing in Cybersecurity*, pp. 295–312, 2023.
24. Lopez, T.I., Twitter-Vaccine Sentiment Analysis, doi:10.2139/ssrn.3942987. PPR:PPR422888.
25. Perikos, I. and Hatzilygeroudis, I., Recognizing emotions in text using ensemble of classifiers. *Eng. Appl. Artif. Intell.*, 51, 191–201, 2016, https://doi.org/10.1016/j.engappai.2016.01.012.
26. Shih, P.-Y. and Chen, C.-P., Computer Science and Engineering Kaohsiung, Taiwan ROC National Chung Kung Univerisity Computer Science and Information Engineering. *IEEE International Conference 011 Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 2756–2760, 2017.
27. Aribisala, B., Olori, O., Owate, P., Emotion Recognition Using Ensemble Bagged Tree Classifier and Electroencephalogram Signals. *J. Res. Rev. Sci.*, 5, 1, 106–111, 2018, https://doi.org/10.36108/jrrslasu/8102/50(0141).
28. Namdev, A., Patni, D., Dhaliwal, B. K., Parihar, S., Telang, S., Rawat, A., Potential Threats and Ethical Risks of Quantum Computing. *Quantum Computing in Cybersecurity*, pp. 335–352, 2023.

29. Mishra, A., Singh, A., Ranjan, P., Ujlayan, A., Emotion classification using ensemble of convolutional neural networks and support vector machine. *2020 7th International Conference on Signal Processing and Integrated Networks, SPIN 2020*, pp. 1006–1010, 2020, https://doi.org/10.1109/SPIN48934.2020.9071399.
30. Youngquist, O., An ensemble neural network for the emotional classification of text. *Proceedings of the 33rd International Florida Artificial Intelligence Research Society Conference, FLAIRS 2020*, pp. 270–275, 2020.
31. Salama, E.S., EI-Khoribi, R.A., Shoman, M.E., Wahby Shalaby, M.A., A 3D-convolutional neural network framework with ensemble learning techniques for multi-modal emotion recognition. *Egypt. Inf. J.*, 22, 2, 167–176, 2021, https://doi.org/lO.1016/j.eij.2020.07.005.
32. Zehra, W., Javed, A.R., Jalil, Z., Khan, H.U., Gadekallu, T.R., Cross corpus multi-lingual speech emotion recognition using ensemble learning. *Complex Intell. Syst.*, 7, 4, 1845–1854, 2021, https://doi.org/10.1007/s40747020-00250-4.

29

Alzheimer Disease Detection Using Machine Learning Techniques

M. Prabavathy[1]*, Paramita Sarkar[2], Abhrendu Bhattacharya[2] and Anil Kumar Behera[3]

[1]Centre for Differently Abled Persons, Bharathidasan University, Tiruchirapalli, Tamil Nadu, India

[2]Department of C.S.E, JIS University, Agarpara, West Bengal, India

[3]Doctorate in National Board, Neurology, DNB Medicine, MBBS, Hyderabad, Telangana, India

Abstract

Alzheimer's disease is a progressive neurological condition that gradually diminishes cognitive function, affects task performance, and disrupts normal behavior. In the absence of regular surveillance, numerous diseases, such as Alzheimer's, remain unnoticed until they have already advanced to a severe stage. Alzheimer's disease is considered a leading cause of death in many countries. The initial phase of Alzheimer's disease is referred to as preclinical sickness, which is followed by stages characterized by mild cognitive and/or behavioral impairment. For the prediction of medical illnesses, a machine learning model is one of the most efficient choices. Most machine learning algorithms are not appropriate for high-dimensional data because to the sparsity problem. As a result, they can only detect Alzheimer's disease using a limited feature space and dimensionality. This study's technique focuses on various machine learning, deep learning, and transfer learning models to enable early diagnosis of Alzheimer's disease.

***Keywords*:** Alzheimer disease detection, machine learning, deep learning, feature selection, classification

**Corresponding author*: cdapraba@bdu.ac.in

Rajesh Kumar Chakrawarti, Ranjana Sikarwar, Sanjaya Kumar Sarangi, Samson Arun Raj Albert Raj, Shweta Gupta, Krishnan Sakthidasan Sankaran and Romil Rawat (eds.) Natural Language Processing for Software Engineering, (443–456) © 2025 Scrivener Publishing LLC

29.1 Introduction

3D scans of the human body, typically taken using a computed tomography (CT) or magnetic resonance imaging (MRI) scanner, are now being used in medical image processing for research, diagnosis, and guiding medical operations like surgery planning. Additionally, medical image processing helps specialists like radiologists, engineers, and physicians to better understand the anatomy of both individual and group patients. The main benefit of medical image processing is that it allows for a thorough evaluation of interior anatomy without intrusive procedures. Better patient treatment outcomes, improved medical equipment and drug delivery systems, and more precise diagnoses are all possible outcomes of building and studying 3D models of relevant anatomy. Chronic diseases, or illnesses that continue for a longer duration, are defined as conditions that last three months or more, according to the US national center for health statistics. In addition to not going away on their own, vaccinations are unable to prevent or cure chronic illnesses. In 1998, 88% of Americans 65 and older were dealing with a chronic condition. Most prevalent chronic illnesses have three main causes: tobacco use, insufficient physical exercise, and poor eating habits. Asthma, cancer, obesity, cardiovascular disease, diabetes, and Alzheimer's disease are just a few examples of the many chronic illnesses that exist. These chronic illnesses are presently affecting the majority of the population. Machine learning (ML) algorithms for early illness prediction are proliferating along with other advances in information and communication technology [1].

For example, Alzheimer's disease (AD) is one of such conditions that, without routine monitoring, would go undetected until it has already developed a significant amount. While advancing age is the leading cause of this cognitive disease, other variables including a serious brain damage can also play a role in the onset of this concerning illness [2]. The capacity to remember and accurately represent information is profoundly affected by the presence of AD. German doctor Dr. Alois Alzheimer found Alzheimer's disease in the brain of a 56-year-old lady after she passed away in 1906. A lady with a history of mental illness and suspicious conduct had her brain altered by Alzheimer's disease. Early onset, moderate, and advanced phases are used to categorize this condition, which is characterized by behavioral and functional dysfunction. Mild Alzheimer's disease patients in the initial stage may seem well on the outside, but they are actually becoming more confused about what's going on in their lives. In contrast, patients in the second stage of Alzheimer's disease require more constant monitoring and care, and those in the last stage are completely reliant on others for their survival [3].

One of the top killers in a lot of nations is Alzheimer's disease. Memory loss is the initial symptom of Alzheimer's disease. Declining physical and cognitive capacities exacerbate patients' already substantial challenges with day-to-day functioning. They will eventually need other people to help them take care of themselves. As a result, caring for people with Alzheimer's has become a major economic and social concern. A correct identification of the illness is crucial for prompt treatment. Mild cognitive impairment (MCI) is the first stage of Alzheimer's disease dementia; those who have it have an increased risk of 10% year for getting the condition. As of now, there is no pharmacological treatment that can restore cognitive function to those who have suffered from Alzheimer's disease or mild cognitive impairment (MCI) [4]. Nonetheless, detecting AD early enough to halt its growth is extremely challenging. Consequently, determining if a patient has AD or MCI becomes an extremely difficult assignment for the clinic. There is ongoing dispute on the primary benefit of early diagnosis of this illness. During the mild dementia stage, loved ones and medical professionals see that a person is experiencing significant memory and cognitive issues that are impacting their ability to carry out daily tasks. This is the hallmark of mild Alzheimer's disease. Over time, the individual and their loved ones may start to notice that something is off. Memory loss, poor judgement leading to poor decisions, lack of initiative and impulsivity, prolonged completion of everyday chores, question repetition, disorientation, roaming, and misplacing items in unusual locations are all possible issues [5].

29.2 Machine Learning Techniques to Detect Alzheimer's Disease

Out of the several medical sickness prediction models, a machine learning model is considered one of the most superior. Due to the sparsity problem, the majority of machine learning algorithms can only detect AD using data and feature spaces that have a minimal number of dimensions. Figure 29.1 depicts it. Several recent suggestions have emerged for predicting AD automatically utilizing advanced clustering and classification techniques that handle high-dimensional data. Current research efforts are focused on developing an enhanced machine learning model to diagnose Alzheimer's disease with greater speed and accuracy [6]. By employing either supervised or unsupervised learning techniques, one may successfully diagnose Alzheimer's disease and forecast its severity. The initial stage in supervised

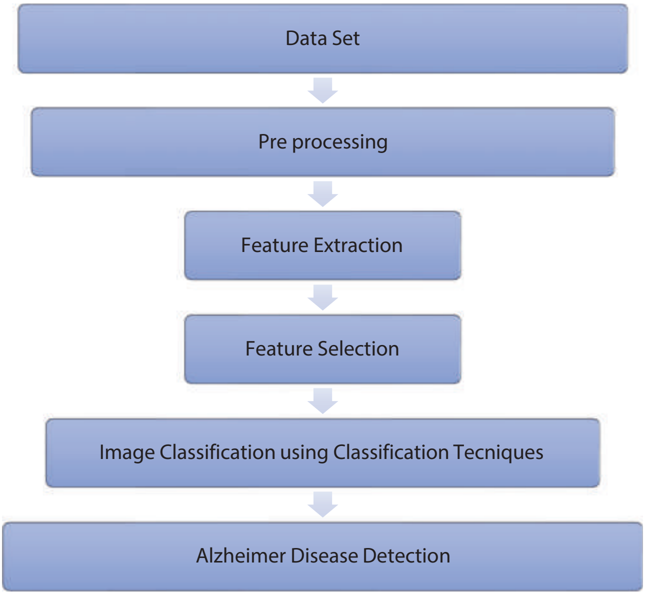

Data Set
Pre processing
Feature Extraction
Feature Selection
Image Classification using Classification Tecniques
Alzheimer Disease Detection

Figure 29.1 Framework to detect Alzheimer's disease using machine learning techniques.

learning involves partitioning the data into a training set and a test set. The training dataset is used to train the classifier, while the test dataset is used to evaluate the performance of the trained classifier model. Typically, the training dataset consists of 70% to 80% of the total data, whereas the testing dataset comprises the remaining 20%. Preprocessing entails examining the dataset for abnormalities, such as missing values or noisy data, once the training dataset has been established. Subsequently, to consistently and distinctly differentiate between the classes, feature extraction is performed. Feature selection is the subsequent stage in improving performance. Ultimately, the process of learning for AD detection is conducted utilizing supervised models such as decision trees (DT), random forests (RF), and support vector machines (SVM). Similarly, the testing dataset is used to assess the performance of the trained classifier.

29.3 Pre-Processing Techniques for Alzheimer's Disease Detection

Prior to utilizing a machine learning model with unprocessed data, it is imperative to do data pre-processing. Semi-automatic and fully-automatic methods have been employed for processing and categorizing AD data on several occasions. Researchers also observed that there are several techniques available for diagnosing AD. Active research is now focused on the discovery of biomarkers that can predict the start of cognitive decline, particularly in the early stages of Alzheimer's disease. Typical in real-world circumstances is the presence of complex, cacophonous, and erratic data. The occurrence of anomalies or inaccuracies in data gathering can be attributed to the rapid increase in data volume and the widespread availability of varied data sources. The capacity to generate dependable models and, consequently, dependable forecasts, relies on the caliber of the data utilized. Hence, it is imperative to process data for best quality. The process of information pre-processing has four main steps: cleaning, integration, transformation, and reduction. Data cleaning is the process of improving the quality of data by filling in missing values and reducing outliers. Integrating several sources of data offers users a consolidated perspective. Data transformation involves modifying the structure, values, or format of data to enhance its use. However, the computing expenses were decreased by implementing data reduction techniques to minimize the dimensions [7].

Data cleaning techniques are employed to rectify mistakes, identify anomalies, and impute missing values, while simultaneously reducing the impact of noise. Filters are essential as they effectively remove noise from the visual data. The Gaussian, mean, median, and bilateral filters are the most often employed filters for image processing, despite the availability of several other options [8]. The process of reducing visual noise is accomplished by utilizing a Gaussian filter. On the other hand, a mean filter, which is essentially a sliding window, calculates the average of all the pixels within the window and assigns it as the central value. A median filter, in contrast to a mean filter, selects the middle value from all the pixels inside the window and assigns it as the central value. Nevertheless, a bilateral filter, which is a non-linear smoothing filter, can effectively decrease noise while preserving edges. Authors propose a method that replaces the intensity of each pixel with a weighted average of intensity values from neighboring pixels. Due to the inherent characteristics of data integrations and transformations, their utilization in medical image processing is limited.

While there are several techniques available for decreasing data, only a limited number are commonly employed to enhance the accuracy of machine learning models. Data reduction solutions utilize the integrity of the original data to generate a smaller dataset while preserving its integrity. Alternatively, one may argue that it serves as a means to decrease the dimensionality of a dataset while preserving the integrity of the data. Data reduction strategies encompass several techniques such as data aggregation into cubes, dimensionality reduction, data compression, elimination of redundancy, data discretization, and creation of a concept hierarchy. Dimensionality reduction techniques are widely employed in machine learning to enhance the accuracy of prediction models for data classification and clustering [9].

29.4 Feature Extraction Techniques for Alzheimer's Disease Detection

Features are essential in image processing. Before extracting features, the sampled image goes through many image pre-processing procedures, including binarization, thresholding, scaling, and normalization. Subsequently, feature extraction approaches are employed to obtain features that may be applied for picture categorization and recognition [10]. The method of feature extraction may be used to divide a large dataset into smaller, more manageable parts. Processing will be simplified as a result. There is a vast array of variables in these extensive databases. Significant computational resources are required to manage these variables. Feature extraction is a process that identifies and selects the most relevant characteristics from large datasets [11]. It involves calculating the number of features in the dataset and creating new features, thereby lowering the amount of data. These properties also uniquely and easily describe the real dataset. Feature extraction is commonly used when raw data is very variable and not useful for machine learning models, and it becomes necessary to convert it into a more suitable format. Potential sources of data that may be analyzed for valuable attributes including text, images, geographic data, web information, audio recordings, and sensor data, among other possibilities. This study attempt incorporates the analysis of both image and audio data. To enhance the diagnosis and classification of Alzheimer's disease, it is customary to search for distinct characteristics in medical imaging [13].

Authors assert that forms play a crucial role in characterizing an object based on its most significant features, hence minimizing the volume of data that has to be stored. The region-based strategy utilizes the method of clustering and assigning labels to all pixels that are associated with objects, considering them as part of a unified region. The image is divided into distinct parts based on certain criteria, and each zone is assigned a unique set of pixels. A contour is a line that connects all the areas along the boundaries of an image that have the same level of brightness [14].

The field of audio analysis is now seeing significant expansion in the use of artificial intelligence and deep learning. Prior to training any statistical or machine learning model, it is necessary to extract the pertinent features of the audio stream. One of the stages involved in the processing of audio signals is the extraction of audio characteristics. The primary emphasis is on the manipulation or alteration of audio signals. It achieves equalization of time-frequency ranges and reduction of undesirable noise by converting both digital and analogue signals [15]. The focus of this topic revolves around the use of computational techniques to manipulate sounds. Audio features refer to descriptions of sounds or audio signals that may be utilized to train statistical or machine learning models for the development of intelligent audio systems. Several audio applications that utilize these characteristics include speech recognition, audio segmentation, source separation, automated music tagging, and audio categorization [16]. Various characteristics capture distinct components of sound that may be heard. Generally, musical signals may be categorized into three levels of abstraction: high-level, medium-level, and low-level. This is grounded in the auditory characteristics. Next, we will discuss the temporal properties of instantaneous, segmental, and global scales. The subsequent musical elements are rhythm, pitch, melody, harmony, timbre (the tonal quality of sound), and beat. Furthermore, the signal domains can encompass both temporal and spectral information, together with automatically retrieved characteristics from deep learning models or features chosen by machine learning models [17].

29.5 Feature Selection Techniques for Diagnosis of Alzheimer's Disease

The use of the labels provided in supervised feature selection aims to improve the ensuing classification process. Unsupervised feature selection aims to improve the clustering of data samples without using the labels of the data samples. There are three primary types of feature selection techniques based on the mechanism used: filter approaches, wrapper approaches, and intrinsic approaches. Filter techniques evaluate properties by examining their statistical characteristics of occurrences. Feature assessment in wrapper approaches relies on the prediction error of a learning algorithm [18]. Wrapper techniques utilize an iterative process of applying a learning algorithm to a specific subset of characteristics and data. This allows them to choose the subsets that yield the most accurate prediction errors. Feature selection is an integral component of the modeling process in intrinsic approaches [19].

Typically, feature selection strategies are organized based on the interaction between machine learning algorithms and feature assessment procedures. The popularity of the feature selection filter approach is driven by the advantages it offers in terms of computing costs and learner independence. Filter techniques for feature selection are constructed without considering any specific machine learning algorithms. The qualities, however, are chosen based on the outcomes of several statistical tests. A set of univariate measurements is employed to rank all the characteristics, and the ones that score highest are chosen. Instead of measuring the error rate, the filter approaches evaluate the feature subset using a surrogate quantity. We chose this quantity because it effectively encompasses the feature set and its usefulness without requiring excessive processing resources [20].

Wrapper techniques employ a novel prediction model to evaluate the feature subsets. The model is trained using individual, newly created subsets and then assessed against a separate control set. A new score is generated for this subgroup by including the actual errors made on a separate set. Wrapper techniques often yield the most optimum feature set for a given model, but they are computationally expensive due to the need to train new models for each subset. To choose the optimal subset, we employ an internal cross-validation approach and a focused classification algorithm to evaluate the true quality of the features. The intrinsic feature selection technique utilizes decision trees to pick features based on metrics. Unsupervised learning depends on correlation to choose features [21].

Significant feature selection procedures include domain knowledge, handling missing data, assessing correlation with the target class, evaluating correlation between features, using principal component analysis, employing forward feature selection, and considering feature significance. A proficient data scientist or machine learning engineer will greatly enhance any case study by effectively identifying characteristics. Real datasets frequently have missing values as a result of data manipulation or oversight in capturing them. Regarding missing values, there are several imputation techniques, albeit their effectiveness is not always guaranteed. Models trained on incomplete data characteristics may not consistently yield higher outcomes. The correlation coefficient is used to assess the degree of connection between each variable and references to the target class. It is a measure of the correlation between the label and attributes of the target class [22]. Two variables are said to be related only if their correlation coefficients are high. Therefore, each modification to a particular characteristic results in changes to associated variables. Matrix factorization preserves variance while reducing the dimensionality of the dataset. It defines a specific group of the most successful characteristics for the machine learning model, while also considering the process of selecting those characteristics. The model's ranking of characteristics is outlined in the feature relevance section. A significance score is awarded to each variable. The filter-based entropy technique was utilized in this work to choose features for AD detection.

29.6 Machine Learning Models Used for Alzheimer's Disease Detection

Machine learning is utilized in several fields such as biomedical systems, digital image processing, knowledge engineering, bioinformatics, and biomedical engineering. The three primary classifications of machine learning methodologies are reinforcement learning, supervised learning, and unsupervised learning. In supervised learning, the training step involves utilizing a large dataset to train the classifier. The testing phase, on the other hand, involves evaluating the performance of the classifier using a different dataset that was not used for training. However, there are instances where the distribution of classes in the training data is uneven. The scarcity of real-time medical data examples hinders the training and testing of classifiers. Common supervised learning models utilized for AD detection encompass Naive Bayes, Support Vector Machines (SVM), logistic regression, K-Nearest Neighbors (KNN), decision trees, neural networks, and random forests [23]. Unsupervised learning does not require data training. Unsupervised learning is a distinct approach from supervised learning, which aims to identify concealed patterns within data for the purpose of grouping related data types [24]. These models can also be utilized in the AD prediction procedure. Common unsupervised learning methods utilized for anomaly detection (AD) encompass hierarchical clustering, C-means clustering, and k-means clustering. Reinforcement learning is a machine learning approach that enables intelligent agents to acquire knowledge by interacting with their surroundings. On the other hand, reinforcement learning is an approach to instructing machine learning that employs rewards for desirable behavior and penalties for subpar performance [25]. An agent that learns by experiential learning possesses the ability to perceive and comprehend its surroundings, execute actions, and acquire knowledge from its errors. The model utilizes reinforcement learning (RL) in combination with domain knowledge and differential equations to simulate the progression of Alzheimer's disease (AD). Specific aspects associated to Alzheimer's disease can be represented by differential equations. To include the missing connections into the model, it is necessary for them to conform to fundamental rules governing brain functionality, such as maximizing cognitive performance while limiting the resources required to sustain cognition [26].

Authors assert that the model may restore the lost connections by optimizing an objective (reward) function that satisfies the aforementioned constraints. The study utilized decision trees, convolutional neural networks, support vector machines (SVMs), and a pre-trained visual geometry group (VGG16) model as supervised models to accurately classify AD classes [27].

Deep learning (DL) is a very innovative advancement in current artificial intelligence (AI) that is also known as hierarchical learning or deep structured learning. Until the 1990s, conventional machine learning methods were used to draw conclusions and make forecasts based on data. However, it had certain flaws. For instance, it depended on manually crafted features and had restrictions because it could only achieve precision at the level of human capability. Within the larger field of artificial intelligence (AI), deep learning (DL) specifically refers to a type of learning algorithm or model that is built upon artificial neural networks (ANN). On the other hand, deep learning has the advantage of extracting features from data during training rather than depending on manually designed feature engineering. Furthermore, Deep Learning has the capability to enhance its predictive and classification abilities by using extensive datasets, state-of-the-art computational power, and creative techniques. The inherent characteristics of DL models, such as their capacity to extract distinctive attributes, non-linear behavior, and extensive parallel processing, have significantly contributed to their efficacy and extensive adoption. Currently, there are several deep learning models accessible, such as convolutional neural networks, recurrent neural networks, long short-term memory networks, and various others. Several more deep learning models have been suggested for computer vision problems following the triumph of AlexNet, a model based on convolutional neural networks (CNNs), including VGGNet, GooglNet, ResNet, DenseNet, and various others. Like in machine learning (ML), the first stage of the framework in deep learning (DL) will include the pre-processing of pictures. To increase the number of samples, augmentation is performed following pre-processing. Subsequently, a deep learning model such as a Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), or similar, would be employed to facilitate the learning process. Finally, the identical model is employed as a classifier to ascertain the different categories of Alzheimer's disease.

Given the alarming worldwide transmission [12] of the illness, it is imperative to now prioritize the screening, isolation, and treatment of AD patients [28]. Regrettably, the extended duration required to get MRI images results in infrequent utilization of testing facilities in hospitals or community health centers. Researchers in this field are dedicating significant time and effort to develop many viable deep transfer learning (DTL) models that utilize Mel spectrogram, derived from audio data gathered from patients, to address this issue. Transfer learning is a technique that can significantly decrease the time and effort required to train and fine-tune a model for a new task. Deep learning (DL) needs a substantially larger quantity of training data compared to standard machine learning methods. Creating extensive, well-annotated medical datasets is difficult and costly, which poses a significant barrier to accomplishing certain crucial domain-specific tasks due to the need for a substantial volume of labeled data. This holds particularly true in the context of medical applications. Despite the diligent work of academics, the traditional deep learning model still needs a substantial amount of computer resources, such as a server equipped with GPU capabilities. DTL, or Domain Transfer Learning, can significantly decrease the training data and time needed for a given job in a target domain. This can be achieved by utilizing a pre-trained model as either a fixed feature extractor or by further fine-tuning.

29.7 Conclusion

Alzheimer's disease, the prevalent kind of dementia in older individuals, is a progressive neurological disorder that imitates many of the brain's activities while gradually diminishing brain capability as a result of neuronal cell demise and tissue deterioration. While there is presently no proven remedy for Alzheimer's disease, there are several medications available that can assist in symptom management, as well as coping strategies to aid with behavioral regulation. Scientists embarked on a quest to identify methods for early detection and prevention of Alzheimer's disease, as an early and precise diagnosis is advantageous for managing the condition. Efficient automated techniques are necessary for the timely detection of Alzheimer's disease. Researchers have proposed several innovative methods to categorize Alzheimer's disease.

To enhance our learning strategies, however, we require a more profound understanding of Alzheimer's research. The current models for forecasting the start of Alzheimer's disease are insufficient.

References

1. Chauhan, D., Singh, C., Rawat, R., & Chouhan, M., Conversational AI Applications in Ed-Tech Industry: An Analysis of Its Impact and Potential in Education. *Conversational Artificial Intelligence*, 411–433, 2024.
2. Durga Prasad Jasti, V., Zamani, A.S., Arumugam, K., Naved, M., Pallathadka, H., Sammy, F., Raghuvanshi, A., Kaliyaperumal, K., Computational Technique Based on Machine Learning and Image Processing for Medical Image Analysis of Breast Cancer Diagnosis. *Secur. Commun. Netw.*, 2022, Article ID 1918379, 7, 2022, https://doi.org/10.1155/2022/1918379.
3. Wen, J., *et al.*, Convolutional neural networks for classifcation of Alzheimer's disease: overview and reproducible evaluation. *Med. Image Anal.*, 63, 101694, 2020.
4. Mishra, A. K., Tyagi, A. K., Dananjayan, S., Rajavat, A., Rawat, H., & Rawat, A., Revolutionizing Government Operations: The Impact of Artificial Intelligence in Public Administration. *Conversational Artificial Intelligence*, 607–634, 2024.
5. Physicians PC, Alzheimer's disease facts and figures. *Alzheimers Dement.*, 16, 3, 391–460, 2020.
6. Yang, Y., Li, X., Wang, P., Xia, Y., Ye, Q., Multi-Source transfer learning via ensemble approach for initial diagnosis of Alzheimer's disease. *IEEE J. Transl. Eng. Health Med.*, 1, 1–10, 2020.
7. Adelina, C., The costs of dementia: advocacy, media, and stigma. *Alzheimer's Dis. Int. World Alzheimer Rep.*, 2019, 100–1, 2019.
8. Suthar, H., Rawat, H., Gayathri, M., & Chidambarathanu, K., Techno-Nationalism and Techno-Globalization: A Perspective from the National Security Act. *Quantum Computing in Cybersecurity*, 137–164, 2023.
9. Noonia, A., Beg, R., Patidar, A., Bawaskar, B., Sharma, S., & Rawat, H., Chatbot vs Intelligent Virtual Assistance (IVA). *Conversational Artificial Intelligence*, 655–673, 2024.
10. Simon, A., Deo, M., Selvam, V., Babu, R., An overview of machine learning and its applications. *Int. J. Electr. Sci. Eng.*, 1, 22–24, 2016.
11. Priyanka and Balwinder, S., An Improvement In Brain Tumor Detection Using Segmentation And Bounding Box. *Int. J. Comput. Sci. Mob. Comput. (IJCSMC)*, 2, 239–246, 2013.

12. Yousuf, M.A. and Nobi, M.N., A new method to remove noise in magnetic resonance and ultrasound images. *J. Sci. Res.*, 3, 1, 81–89, 2010.
13. Ramalakshmi, C. and Chandran, A.J., Automatic brain tumor detection in MR images using neural network based classification. *Biom. Bioinf.*, 5, 6, 221–225, 2013.
14. George, E.B. and Karnan, M., MRI Brain Image enhancement using filtering techniques. *Int. J. Comput. Sci. Eng. Technol. (IJCSET)*, 3, 399–403, 2012.
15. Manjón, J.V., Carbonell-Caballero, J., Lull, J.J., García-Martí, G., MartíBonmatí, L., Robles, M., MRI denoising using non-local means. *Med. Image Anal.*, 12, 4, 514–523, 2008.
16. Sakthivel, K., Jayanthiladevi, A., Kavitha, C., Automatic detection of lung cancer nodules by employing intelligent fuzzy c means and support vector machine. *Biomed. Res.*, 27, 123–127, 2016.
17. Seixas, F.L., Zadrozny, B., Laks, J., Conci, A., Saade, D.C.M., A Bayesian network decision model for supporting the diagnosis of dementia, Alzheimer's disease and mild cognitive impairment. *Comput. Biol. Med.*, 51, 140–158, 2014.
18. Chauhan, D., Singh, C., Rawat, R., & Dhawan, M., Evaluating the Performance of Conversational AI Tools: A Comparative Analysis. *Conversational Artificial Intelligence*, 385-409, 2024.
19. Vemuri, P., Gunter, J.L., Senjem, M.L., Whitwell, J.L., Kantarci, K., Knopman, D.S., Jack, C.R., Alzheimer's disease diagnosis in individual subjects using structural MR images: validation studies. *Neuroimage*, 39, 3, 1186–1197, 2008.
20. Gad, A.R., Hassan, N.H., Seoud, R.A.A., Nassef, T.M., Automatic Machine Learning Classification of Alzheimer's Disease Based on Selected Slices from 3D Magnetic Resonance Imagining. *Age*, 67, 10–15, 2016.
21. Siddiqui, M.F., Mujtaba, G., Reza, A.W., Shuib, L., Multi-Class Disease Classification in Brain MRIs Using a Computer-Aided Diagnostic System. *Symmetry*, 9, 3, 1–37, 2017.
22. Beheshti, I. and Demirel, H., Feature-ranking-based Alzheimer's disease classification from structural MRI. *Magn. Reson. Imaging*, 34, 3, 252–263, 2016.
23. Xiao, Z., Ding, Y., Lan, T., Zhang, C., Luo, C., Qin, Z., Brain MR Image Classification for Alzheimer's Disease Diagnosis Based on Multi feature Fusion. *Comput. Math. Methods Med.*, 1, 1–14, 2017.
24. Bhateja, V., Rastogi, K., Verma, A., Malhotra, C., A non-iterative adaptive median filter for image denoising. *International Conference on Signal Processing and Integrated Networks (SPIN)*, pp. 113–118, 2014.
25. Kaplan, I., Applying the Haar wavelet transform to time series information, pp. 1–176, 2001.
26. Zhao, W. and Wang, J., A new method of the forest dynamic inspection color image sharpening process. *2010 3rd International Conference on Advanced*

Computer Theory and Engineering (ICACTE), pp. V4–211-V4-214, 2010, doi: 10.1109/ICACTE.2010.5579715.

27. Rawat, R., Mahor, V., Rawat, A., Garg, B., & Telang, S. Digital transformation of cyber crime for chip-enabled hacking. In *Handbook of research on advancing cybersecurity for digital transformation* (pp. 227-243), IGI Global, 2021.
28. Ibrahim, I. and Abdulazeez, A., The Role of Machine Learning Algorithms for Diagnosing Diseases. *J. Appl. Sci. Technol. Trends*, 2, 01, 10–19, 2021, Available: 10.38094/jastt20179.

30

Netnographic Literature Review and Research Methodology for Maritime Business and Potential Cyber Threats

Hitesh Rawat[1]*, Anjali Rawat[2] and Romil Rawat[3]

[1]Department of Business Management and Economics, University of Extremadura, Badajoz, Spain
[2]Department of Computer and Communication Technology, University of Extremadura, Badajoz, Spain
[3]Department of Computer Science and Engineering, Shri Vaishnav Vidyapeeth Vishwavidyalaya, Indore, India

Abstract

Worldwide apprehension over transnational organized crime at sea is growing. People are focusing on cyberthreats and crime-related convergences as a result of this uncertainty, which is impeding the development of a unified foreign policy response (ocean and maritime criminology and protection). The work focuses on the comprehension gap by conveying a compositional conceptualization and examinination of cross-border organized crime in the Pacific. The proposed work represents information, including observations of cyberthreats to the marine industry. A netnographic analysis is conducted on the OSN channels, where vulnerabilities are analyzed and disseminated. The method involves conducting a netnographic study on the digital platform, and entails making long-term descriptive observations and interpretations of online social networks and platforms. The framework for a promising investigation into the various types of oceanic crime committed by malicious actors to map their unique characteristics is also discussed in the work.

***Keywords*:** Ocean crime, maritime, cyber security, illicit trafficking, cyber attack, business threat

**Corresponding author*: hrawat88@gmail.com

Rajesh Kumar Chakrawarti, Ranjana Sikarwar, Sanjaya Kumar Sarangi, Samson Arun Raj Albert Raj, Shweta Gupta, Krishnan Sakthidasan Sankaran and Romil Rawat (eds.) Natural Language Processing for Software Engineering, (457–474) © 2025 Scrivener Publishing LLC

30.1 Introduction

A qualitative research approach called ethnographic [34] research uses in-person observation to examine individuals in their natural settings. It is a fundamental approach in feminist geography, development studies, cultural ecology, and the social and behavioral sciences and is frequently employed in these fields. This may involve researching all facets of human communication, including tone of voice, body language, and netnography—all of which are research methodologies.

Netnography [34] is a research approach that focusses on online communities and digital cultures. It is characterized as participant-observational research based on OSN that incorporates methods of data gathering such as in-depth interviews conducted on virtual platforms and participant observation.

Ocean piracy, human trafficking [22, 23], drug trafficking [24, 25], weapons trafficking, and waste trafficking by the ocean, as well as environmental crimes, are all becoming important aspects of ocean monitoring [18] and governance, oceanic security, and law enforcement. Such violations take various forms throughout the world's oceans, causing political, social, national, and economic disruptions, ranging from the effects on foreign shipping to Dedicated channels and timely web links created for clients and actors involved in illicit activities who like to trade across nations [19–22].

At the highest levels of foreign policymaking, oceanic piracy is getting more recognition. In February 2019, the United Nations Security Discussion on the subject was titled "Transnational Organised Crime in the Ocean as a Challenge to International Peace and [1] Security." The topic of debate was the effects of the oceanic dark web trade and crime. One of the reasons for this was the widespread misunderstanding and illegal practices in the diverse nature of oceanic cybercrime [17] and [18]. We begin with a review of current concepts of cross-border organized crime in the ocean related to international law, especially the UN Convention on Transnational Organised Crime (UNTOC) [3]. Perhaps such offences, such as piracy, are politically prioritized and reasonably well researched, whereas some are not. Others, such as cocaine abuse, are well recorded and known in some areas but not in others. Conducting more in-depth research on the content and scope of available evidence on any oceanic crime is thus a critical challenge for ocean criminology and police responses.

Business disruptions [2] and cyber events are sometimes referred to as assaults or significant problems. The most troublesome attacks might actually be those that are most localized. Email phishing tactics do indeed still rule the world of cybercrime [3]. While concerns like WannaCry, cryptocurrency mining, and state-sponsored assaults [4] are in the news, lower-level cyberattacks are having a much greater impact, and cargo is by no means immune. A phishing campaign is when a cybercriminal uses a compromised email account to deceive victims into disclosing sensitive or confidential data. Phishing scams and corporate email hack assaults aim to scam companies at all levels [5]. The shipboard [6] equipment of today is well integrated, yet poorly protected. These dangers are high. Key systems, particularly navigation devices, might be jeopardized if attacked or if a virus was unintentionally downloaded, as ships rely more and more on automation [7] and remote monitoring.

Social engineering [8] is combined with phishing, since the originator attempts to make the receiver believe they are someone, they are not. But effective phishing [9] attempts can also be attributed to inadequate online safety. It describes the procedures and actions computer users take to maintain system health and enhance internet privacy. A vessel that had been the victim of a cyberattack has been sailed on by over half (49%) of commercial seamen. There are no signs that cyberpiracy will decrease as a result of improving technology and rising freight values. In a marine poll, 45% [13] of participants revealed that their only cyber security measure was a simple firewall. Figure 30.1 shows the crime increase rate in marine territory.

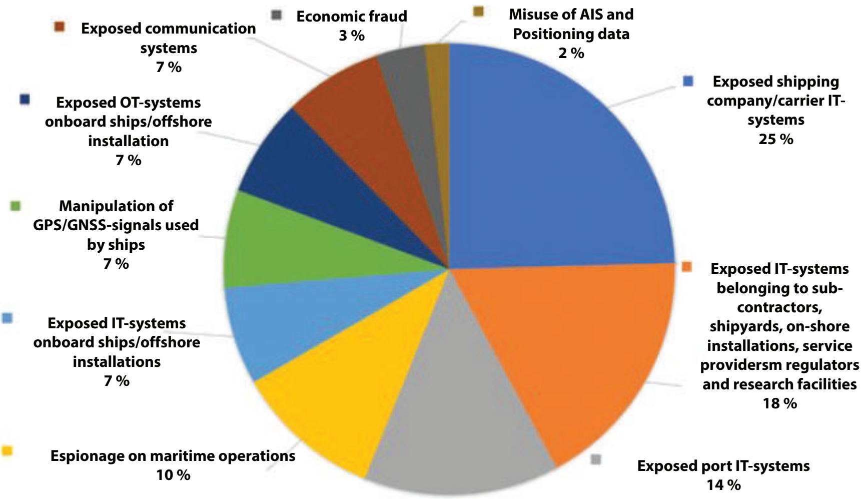


Figure 30.1 Crime increase rate at marine territory [23, 28].

The International Maritime Organization (IMO) [11, 12], which advises that cyber security is a crucial component of risk mitigation, is at odds with this. Internet piracy [14] lacks the passion and excitement of traditional piracy. Because modern piracy is carried out behind locked doors hundreds of kilometers [15] away, faster ships and greater weapons will not be able to prevent cyberattacks or keep sailors safe. The maritime industry [16, 17] is confronting an additional challenge as a result of the fast digitization of its operations and information, as is the case for many other businesses [18] throughout the world. This digital [19] transition has numerous implications and advantages, but if the proper precautions are not taken to maintain the security of ships, it might easily be used by bad actors.

The rest of the paper is organized as follows: Section 30.2. Shows the Criminal Flows Framework, Section 30.3. Describes Oceanic Crime Exchange and categorization, Section 30.4. It is about fishing crimes, mobility crimes, and finally, Section 30.5. This concludes the paper with a discussion in Section 30.6.

30.2 Criminal Flows Framework

The division is classified into four categories, namely: (1) Crime Mutation; (2) Forced Crimes; (3) Military Activities; and (4) Atmosphere-Disturbing. The categorization is further discussed and presented in Table 30.1: Oceanic Crimes Categorization.

The majority of global smuggling and trafficking activities are tied to oceanic space. Different types of illegal flows in the ocean can be differentiated based on what is transported. Human trafficking, organ trafficking rackets [3, 15, 16] and piracy are serious problems in many oceanic countries, with refugees often forced to fly in precarious boats. Such practices are frequently promoted by sophisticated criminal networks [12]. Human smuggling and trafficking is done for forced labor, prostitution [12, 13], arms smuggling [14, 17] for war zones, weapons of mass destruction and testing, kidnapping, slavery, economic and home instability, narcotic trafficking, and organ trafficking. The ocean is an open and well-known illegal activity hub for loading anything into shipping containers for concealment. Higher rates of overdose, HIV/AIDS, and domestic abuse are all consequences of the opioid trade in coastal areas and beyond [14]. Heroin, opium, hemp, and methamphetamine are among the narcotics trafficked by Ocean [13]. Arm trafficking [28, 29] has the potential to destabilize nations by causing hidden wars among nationals. The movement of war

Table 30.1 Oceanic crimes categorization [3, 4].

Sub categories	Crime mutation	Forced crimes	Military activities	Atmosphere disturbing
Ocean Relation and Area of Vulnerability.	Within the Ocean Territory	Across the Ocean	Under and across the Ocean	Under the Ocean
Crime Classes.	Ship and Cargo capturing, Drug trafficking, National Security Threat, Illicit Trade, Tax Evasion	Fisheries Crimes, Blackmailing, Organ Trafficking, Unauthorized Entry	Disastrous Weapons Showcasing and Testing.	Plants and Animal Trafficking, industrial Waste Pollution and Discharging, Illicit Mining, Violence at Ship Attack by Missile
Affecting Objects and Organizations.	Cargo ,Marine and Shipping Industry	Healthcare, Food Industry, Destroying Cultural Heritage	Border Enhancement and Capturing.	Biodiversity Disturbance.
Crime Categories.	Slavery and Illegal Entry into Countries	Child and Human Trafficking, Ancient Antiques Trafficking.	International Law and Treaty Dissolution	Natural Resources Disturbance, myths about missing of ships
Reasons of Crime and Inability of Security agencies to Track and Stop the Vulnerable Events.	Unchecked Power and Support by Terrorist Groups, Trained Pirate Attacks	Passing from Criminal Zones, To hide Militant from International Security Agencies by Continuously moving over Ocean	Supremacy and Superpower Competency Across the Global.	Coastal Locations Limitation for Checked and Controlled Law Enforcement
Automated Crime	Control System Hacking by Unidentified Hackers (Cyber Criminals).	Malware Injection and Security Penetration.	Secure Location Coordinates Tracing and Automated Attack Installation.	Online Bidding at illicit Forums and Transaction.

groups and materials or precursors could jeopardize national and international security. Regulated or banned items such as counterfeit merchandise, antiquities, animals, hardwood lumber, or waste, among others, can be trafficked at Ocean [2]. Illicit trade [30–32] in goods will jeopardize environmental efforts, encourage ecosystem loss, and endanger biodiversity [16]. Finally, licit products like gold, charcoal, gasoline, tobacco, and sugar can be diverted to circumvent taxes, customs duties, and foreign sanctions [11]. Avoiding taxes and customs fees increases the amount of money associated with illegal trade and activities.

30.3 Oceanic Crime Exchange and Categorization

The research suggests that more research is needed to better grasp how various actors and organizations participating in the battle against oceanic crime [5] exchange knowledge with one another and with the general public. This is partly due to questions over accountability, confidence, and the availability of effective data-sharing networks. It does, however, raise methodological concerns, such as the degree to which data can be compared or obtained using standard categories and meanings. The relationship between how oceanic crimes are conceptualized, interpreted, security-focused responses, law enforcement and social justice interventions, underlying causes of crime, community growth, or capacity-building programs is unlikely to be successful, and attention must be given to how different solutions can work together.

The transport of goods and oceanic shipping are the targets of crimes against mobility. Crimes occur on the water or in close proximity to the shore. Movements in the ocean are the subject of crimes against mobility. The ship and the port are the primary targets of damage, with the harmed items being not only vehicles, goods, and their crews, and supply chain destruction is a major offense. Different criminal movements exist. Across the ocean, illegal acts are taking place. Although the damage is done on shore, the ocean provides an outlet for criminals. These offenses have a negative impact on families and cultures. The main crime in this group is smuggling in all sorts of goods. Environmental violations occur in the ocean and are directed at the oceanic environment, including natural resources, installations, and artifacts. The main practices included in this group are illegal mineral trafficking and offenses against oceanic resources. Table 30.1 shows a summary of each division.

A. **Crime Mutation**

A major type of oceanic crime involves crimes in which the ocean is mainly used as a medium for illegal activity rather than the primary location of the crime. The high oceans still have looser and more ambiguous regulatory regimes than those jurisdictions of individual nations, reducing the chance of being apprehended and prosecuted for drug trafficking.

There are frequently associated violent crimes with fishing or recreational boating in general with the courage and desire to use violence with guns or light firearms in war communities with gun cultures. For example, an illegal vessel may engage in lawful fishing, fishery violations, different types of trafficking, or even piracy [9]. Because of these power synergies, oceanic law enforcement can find it difficult to discriminate between legal users of the ocean and those involved in oceanic crime. This may be because such roads, such as those in Djibouti, Somalia, and Yemen, are of vital geo-economic significance for oceanic traffic and are targeted by criminals to commit crime [7]. These areas will serve as chokepoints for oceanic traffic, causing ships to slow down to navigate safely and increase their susceptibility to attack [16].

Criminals often use informal trade networks to transport various illegal goods, such as weapons, people, and drugs. The Western Indian Ocean's show trading network, for example, has been frequently connected to smuggling activities [7]. Overlap may occur through common modes of criminal enterprise or commercial activities that aid in the commission of crimes. For illegal markets, banking, tax fraud, money exchanges, and laundering networks are among them. Traditional activities like bribery or forgery, slave labor, or unethical bribes to ministers, private workers, and criminals may participate in the legal economy in a variety of ways. Table 30.2 depicts the increase in crime rate [26, 27] over the previous four years.

Table 30.3 shows the marine threat generated targeting [23, 26] (port, ocean business, military activities, and trafficking). Table 30.4 represents the Comparative Work Analysis.

Figure 30.2 and Figure 30.3 shows the comparative study of cybercrime that penetrated the security infrastructure of the ocean in Year 2022 and in 2024 respectively, and caused millions of losses towards security organizations and trade.

Table 30.2 Increased in crime rate chart [7, 8].

Ref.	Crime categorization/ identified cases(year) in percentage	2018(in %)	2019(in %)	2020(in %)	2021(in %)	2022(in %)	2024(in
[2, 8]	Ocean Relation and Area of Vulnerability.	18	25	48	51	59	63
[3, 9]	Crime Classes	36	41	46	57	63	68
[4, 10]	Affecting Objects and Organizations.	43	56	67	74	77	82
[5, 11]	Crime Categories.	53	68	77	82	87	91
[6, 12]	Reasons of Crime and Inability of Security agencies to Track and Stop the Vulnerable Events.	44	51	67	87	91	93
[7, 13]	Automated Crime	59	69	78	82	87	89

Table 30.3 Generated marine threats [32–34].

Crime	2024 (Crime-increased rate) (in %)
Capturing , Drug trafficking , National Security Threat, Illicit Trade, Tax Evasion	83
Fisheries Crimes, Blackmailing, Organ Trafficking, Unauthorized Entry	77
Healthcare, Food Industry, Destroying Cultural Heritage	91
Child and Human Trafficking, Ancient Antiques Trafficking, Slavery and Illegal Entry into Countries	95
Passing from Criminal Zones ,To hide Militant from International Security Agencies by Continuously moving over Ocean	75
Unchecked Power and Support by Terrorist Groups, Trained Pirate Attacks	77
Malware Injection and Security Penetration, Control System Hacking by Unidentified Hackers (illicit agents and Criminals).	87

Table 30.4 Comparative work analysis.

Reference	Approach
[22]	Cyberthreats to the water transport industry are explored.
[23]	Based on the pertinent area of international law, the paper elaborates on the legality of eavesdropping on underwater communications cable networks.
[24]	This paper contrasts the development of cyberspace as a man-made realm with the history of the creation of maritime laws, conventions, customs, and standards of conduct.
[25]	Cyberthreat case patterns involving international crime are addressed.
[26]	The topic of the discussion is the economics, safety, and security of the oceans for the ensuing decades, as well as current civilian and military foresight reports, studies, and articles from five continents.
[27]	For new paths spanning digital and green criminology, academics are encouraged to create reflective interactions with post-anthropocentric criticisms.
Proposed Work	Structural and conceptualization study of cross-border organized crime in the ocean are discussed

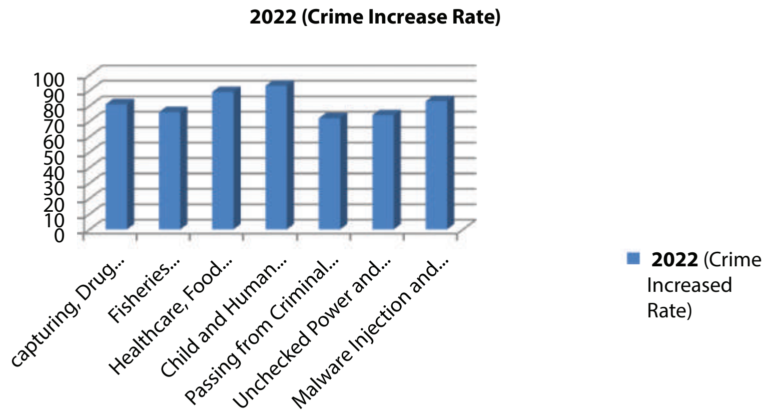


Figure 30.2 Crime increase rate at marine territory (2022).

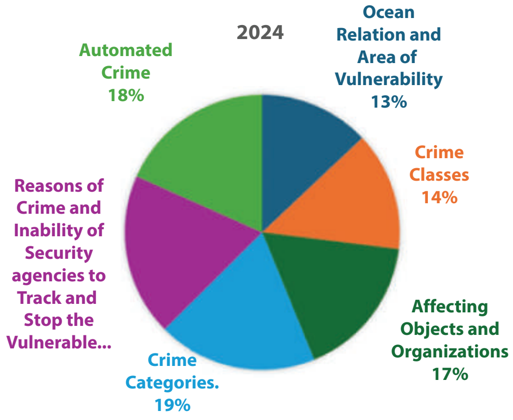


Figure 30.3 Crime increase rate at marine territory (2024).

B. **Atmosphere Disturbing Crime**

Environmental offenses apply to actions that have a direct negative impact on the aquatic environment, with humans being only a secondary victim. Such offenses happen in water, in the context of destruction or depletion of the ocean's fauna, wealth, cultural history, and infrastructure. Anthropocene offenses occur in the form of human actions that communicate and interfere with the ocean's ecosystem [14]. This suggests a greater view of "the climate," acknowledging that it is impossible to separate nature and society in the Anthropocene [3]. Human-made artifacts, shipwrecks, pipelines, wind turbines, and oil platforms are all so intertwined with nature that it's impossible to set them apart. Such environmental violations are often significant and occur with differing levels of organization. They primarily consist of financial gain-oriented breaches of environmental regulations, such as intentional contamination and waste disposal at Ocean, ship ballast water discharge and unchecked breakage operations [8], or the illegal exploitation of natural resources at Ocean [2].

Offshore platforms and vital facilities, such as pipelines and underwater data cables, are all used in the oceanic world. They are vulnerable to illegal activity, such as accidental harm or intentional threat of criminal intent [9]. Crimes against cultural heritage are also included here for related reasons. Treasure searching, pillaging antiquities, and war graves, such as the plundering of sunken warships for scrapping purposes, are examples. Such offenses can have an environmental effect, releasing pesticides or fuels within water bodies that have new biodiversity hotspots [4].

Offenses in the ocean have a wide range of pathological consequences, and illegal fishing puts species in jeopardy. Coral reefs and other underwater habitats are harmed by destructive fishing techniques. Waste disposal and other polluting practices may have disastrous consequences for wildlife and ocean health. Such activities will reduce the productivity and profitability of fishing grounds for legal fishermen, their livelihoods, and fragile coastal areas [6].

Figure 30.4 describes the marine activities [15] that have increased so much in recent years.

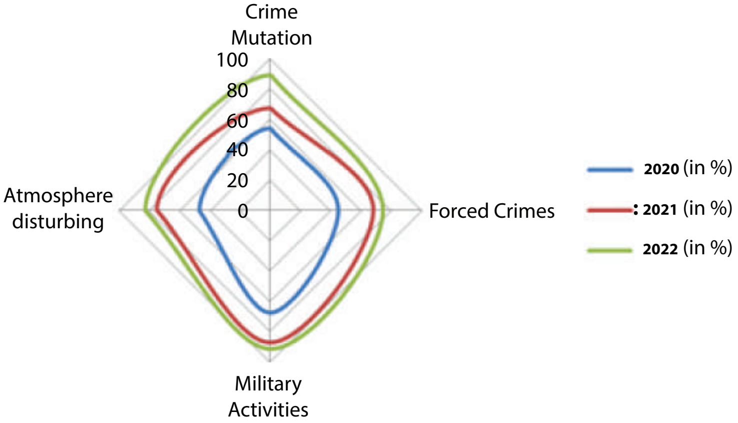


Figure 30.4 Increased marine activities.

C. **Forced Crimes**

Oceanic crime is adaptable and complex. Criminals that are involved in one kind of crime may also be involved in another at the same time or move from one to the other [8]. Motivations resulting from unintended outcomes are the reasons for the transformation and adaptation of oceanic crime and the criminal market, motivated by risk and reward equations, according to criminologists [11]. When the chances of committing a specific type of crime in a specific location become too high, law enforcement action, defensive strategies, criminal justice practices, and offenders are more likely to turn to fewer countermeasures and more favorable risk-reward balances [12]. Piracy coast of Somalia is a good example of the establishment of an effective court prison system, incarceration of pirate suspects for prosecution, oceanic patrolling, and usage of protective armed guards on submarines [13]. It's assumed that the major pirate organizational frameworks are still in place, and that their leaders are still on the loose [4]. Since these organizations are no longer as heavily involved in piracy, they have also broadened their scope to include the transfer of illegal piracy proceeds into legal businesses on the Somali coast, and their networks expertise adapted skills, such as weapons and people trafficking [7].

Countermeasure-driven adaptation often coexists with opportunity-driven motives. When such crimes become more dangerous, new opportunities can arise as a result of geopolitics or criminal creativity to replace them. Changing geopolitical circumstances frequently constrain such opportunities. A rise in war or insecurity in a given nation or area, for example, may result in new armament demands or an influx of refugees fleeing violence or deprivation. For example, the outbreak of Yemen's war in 2015 provided new possibilities for trafficking operations both to and from the country through [3].

Similarly, the destabilization of more developed roads by Russia and the Balkans (Western Indian Ocean) resulted in the advent of a crucial outlet for the transport of Afghani heroin (Northern and Balkan routes). Since 2014, European sanctions against Russia have increased border protection in Eastern Europe and blocked trafficking along the northern pathway. The Balkan route was disrupted by significantly improved border control between Turkey and Iran as a result of the Syrian conflict [5]. On the other hand, global wars, such as the one in Syria, will make operations more complex. Such conditions pose risks to those operating in or passing through them [7].

D. **Military Activities**

Interconnections between oceanic crime and the broader security system include conflict, chaos, and state vulnerability, among others, but it's also worth noting that oceanic crime is linked to geopolitics and terrorism. In the South China Sea [9], illicit fishing interacts with inter-state tensions and geopolitical rivalry. Solid evidence indicates that some of these organizations may participate in oceanic crime to fund their operations. Smuggling of sugar and charcoal in Somalia is linked to the Al Shabab group [16], as is kidnapping and killing for ransom in the Sulu and Celebes Seas, which is linked to the Abu Sayyaf group [5].

30.4 Fisheries Crimes and Mobility Crimes

Following that, Landing endangered animals or using prohibited methods (cyanide or dynamite fishing) are examples of illegal fishing linked to a slew of fishing industry issues, including forgery of documents or tax evasion, trafficking of crew members on board ships, forced labor and slavery [12], collectively termed as "fisheries crimes" [13]. Illegal fishing is expected to cost the fishing industry up to USD 23.5 billion a year, with other fishing-related offenses causing far more financial harm [13]. Artisanal fishermen in restricted areas on a habitual basis may engage in illegal fishing at a low level. However, it is frequently a highly complex and organized crime, involving cross-border criminal networks operated by shell corporations and using several flags of convenience.

The UNCLOS [3] concept has been criticized for its inflexibility and failure to catch certain forms of criminality, particularly referred to as contemporary oceanic piracy [1]. It is a complex phenomenon encompassing a wide range of business structures, which are regionally unique and classified as crime in sovereign waters. Crime in the Gulf of Guinea is focused on the region's oil economy and occurs in the territorial waters of countries like Nigeria [13]. Stealing of cargo or kidnapping of crews is a common crime in the Malacca Straits and Southeast Asia. The bulk of such events, as in the Gulf of Guinea, occur in territorial waters [4]. The guidelines for preventing crime in the maritime industry use a broad term that includes "the use of violence against the ship using digital autonomy, its passengers, crew, or cargo, or any attempt to use violence" [5]. Theft at sea is the result of fraud, and offenses against vehicles are possible at port facilities [9]. Data breaches caused by malware (ransomware) are an emerging criminal practice and have proven difficult to quantify due to concerns about reputational harm for targets [12]. They are frequently well-organized, costing the oceanic industry a lot of money in terms of disembarkation-repatriation conditions and possible penalties for ship owners [13].

30.5 Conclusion

The concept of cross-border organized digital crime in the ocean has made it to the United Nations Security Council and is constantly being designed for discussion on oceanic security and governance. This paper has taken the argument a step further by outlining the essential categories. This lays the groundwork for understanding not just the differences between different forms of crime but also how they connect and bind together. The UNODC and Interpol emerged as the main global organizations tackling oceanic crime by forming governance structures focusing on crimes against the mobility of shipping and port regulations. Border and customs agencies are all focused on criminal movements. Environmental offenses are handled by environmental bodies such as the UNEP (United Nations Environment Program) and the FAO (Food and Agriculture Organization), which lack experience in dealing with violence. Crimes against utilities get very little attention. The international approach to oceanic crimes would have to include rethinking mechanisms and tackling structural instability on both a regional and global basis.

30.6 Discussion

The world community is becoming increasingly concerned about cross-border organized crime in the ocean. This ambiguity has impeded a coordinated foreign policy response by forcing individuals to focus on specific digital risks that intersect with other crimes (ocean and maritime criminology and protection). We fill this gap in our knowledge by presenting a structural conceptualization and study of cross-border organized crime in this article.

References

1. Tachie-Menson, E.A., Investigating the Intersection of Maritime and Cyber Crime in the Gulf of Guinea. *Sci. Mil. S. Afr. J. Mil. Stud.*, *51*, 3, 89–112, 2023.
2. Manhas, N.S., Maritime Security Cooperation Between India and ASEAN, in: *India and ASEAN in the Indo Pacific: Pathways and Perils*, pp. 43–64, Springer Nature Singapore, Singapore, 2024.
3. Rawat, R., Mahor, V., Rawat, A., Garg, B., Telang, S., Digital Transformation of Cyber Crime for Chip-Enabled Hacking, in: *Handbook of Research on*

Advancing Cybersecurity for Digital Transformation, pp. 227–243, IGI Global, USA, 2021.

4. Vrancken, P., State jurisdiction to investigate and try fisheries crime at sea. *Mar. Policy*, 105, 129–139, 2019.
5. Guilfoyle, D., Prosecuting pirates: a critical evaluation of the options. *J. Int. Crim. Justice*, 10, 4, 767–796, 2012.
6. Ford, J.H. and Wilcox, C., Shedding light on the dark side of maritime trade – a new approach for identifying countries as flags of convenience. *Mar. Policy*, 99, 298–303, 2019.
7. Miller, D.D. and Sumaila, U.R., Flag use behaviour and IUU activity within the international fishing fleet: definitions and identifying areas of concern. *Mar. Policy*, 44, 204–211, 2013.
8. Makarenko, T., The crime-terror continuum: tracking the interplay between transnational organised crime and terrorism. *Glob. Crime*, 6, 1, 129–145, 2004.
9. Ekici, B. and Coban, A., Afghan heroin and Turkey: ramifications of an international security threat. *Turk. Stud.*, 15, 2, 341–364, 2014.
10. Gunavathi, R. and Bharathi, K.M., Cybercrimes in the Associated World, in: *Artificial Intelligence for Cyber Defense and Smart Policing*, pp. 21–31, Chapman and Hall/CRC, USA, 2023.
11. Wyatt, T., The local context of wildlife trafficking: the heathrow animal reception centre, in: *Emerging Issues in Green Criminology: Exploring Power, Justice, and Harm*, D. Westerhuis, R. Walters, T. Wyatt (Eds.), pp. 108–126, Palgrave Macmillan, Basingstoke, 2013.
12. Wyatt, T., *Wildlife Trafficking: A Deconstruction of the Crime, the Victims, and the Offenders*, Palgrave Macmillan, Basingstoke, 2013.
13. Gikonyo, C., The Jeddah amendment and the fight against wildlife trafficking. *Crim. Law Forum*, 30, 2, 181–200, 2019.
14. Rawat, R., Mahor, V., Chirgaiya, S., Rathore, A.S., Applications of Social Network Analysis to Managing the Investigation of Suspicious Activities in Social Media Platforms, in: *Advances in Cybersecurity Management*, pp. 315–335, Springer, Cham, 2021.
15. Rawat, R., Mahor, V., Chirgaiya, S., Shaw, R.N., Ghosh, A., Sentiment Analysis at Online Social Network for Cyber-Malicious Post Reviews Using Machine Learning Techniques, in: *Computationally Intelligent Systems and their Applications*, pp. 113–130, 2021.
16. Rasmussen, J., Sweet Secrets: Sugar Smuggling and State Formation in the Kenya-Somalia Borderlands, 2017, DIIS Working Paper 11.
17. UNODC, United Nations Office on Drugs and Crime, Transnational Organized Crime in the Fishing Industry, 2011, Available at, http://www.unodc.org/documents/human-trafficking/Issue_Paper_-_TOC_in_the_Fishing_Industry.pdf.
18. Chalk, P., *The Maritime Dimension of International Security: Terrorism, Piracy, and Challenges for the United States*, RAND Corporation, USA, 2008.

19. Maritime Cyber-Attacks Increase by 900% in Three Years, Available online: https://www.marineinsight.com/shipping-news/maritime-cyber-attacks-increase-by-900-in-three-years/# (accessed on 20 July 2020).
20. https://www.unodc.org/unodc/en/piracy/index.html
21. Senarak, C., Port cybersecurity and threat: A structural model for prevention and policy development. *Asian J. Shipp. Logist.*, *37*, 1, 20–36, 2021.
22. Caponi, S.L. and Belmont, K.B., Maritime cybersecurity: a growing threat goes unanswered. *Intell. Prop. Tech. L. J.*, *27*, 1, 16, 2015.
23. [22] Alekseenkov, A., Klyuchnikova, D., Dedova, N., Sokolov, S., Cyberattacks in the Water Transport Industry: Types and Diversity, in: *International Scientific Siberian Transport Forum TransSiberia-2021: Volume 2*, 2022, March, Springer International Publishing, Cham, pp. 1532–1540.
24. Botua, H.L., Dewi, C.T.I., Siswandi, R.A.G.C., WIRETAPPING ON SUBMARINE COMMUNICATIONS CABLE: QUESTIONING ITS LEGALITY AMIDST LONG STANDING PRACTICE. *Padjadjaran J. Int. Law*, *6*, 1, 20–42, 2022.
25. Howard, L.C.T.D. and da Cruz, J.D.A., Like the Sea, So Cyberspace: A Brief Exploration of Establishing Cyberspace Norms through a Maritime Lens. *J. Adv. Mil. Stud.*, *13*, 2, 142–153, 2022.
26. Rawat, R., Logical concept mapping and social media analytics relating to cyber criminal activities for ontology creation. *Int. J. Inf. Technol.*, *15*, 2, 893–903, 2023.
27. Fita, G.A., Ismira, A., Khaldun, R.I., Fatra, D., Patterns of Transnational Crime in The Border of Sulawesi Sea-Sulu Sea and Threats to Indonesia's Maritime. *Resolusi: J. Sos. Polit.*, *5*, 2, 133–142, 2022.
28. Lauro, A. and Corrêa, C.R., Futures for the Maritime Domain: Signs and Trends That Shape Scenarios, in: *Power and the Maritime Domain*, pp. 286–301, Routledge, USA, 2023.
29. Rawat, R., Gupta, S., Sivaranjani, S., Cu, O.K., Kuliha, M., Sankaran, K.S., Malevolent Information Crawling Mechanism for Forming Structured Illegal Organisations in Hidden Networks. *Int. J. Cyber Warf. Terror. (IJCWT)*, *12*, 1, 1–14, 2022.
30. Bedford, L., Mann, M., Foth, M., Walters, R., A post-capitalocentric critique of digital technology and environmental harm: New directions at the intersection of digital and green criminology. *Int. J. Crime Justice Soc. Democr.*, *11*, 1, 167–181, 2022.
31. Rawat, R., Chakrawarti, R.K., Vyas, P., Gonzáles, J.L.A., Sikarwar, R., Bhardwaj, R., Intelligent Fog Computing Surveillance System for Crime and Vulnerability Identification and Tracing. *Int. J. Inf. Secur. Priv. (IJISP)*, *17*, 1, 1–25, 2023.
32. Meland, P.H., Bernsmed, K., Wille, E., Rødseth, Ø.J., Nesheim, D.A., A retrospective analysis of maritime cyber security incidents. *TransNav: Int. J. Mar. Navig. Saf. Sea Transp.*, *15*, 123–139, 2021.

33. Rawat, R., Oki, O.A., Sankaran, K.S., Olasupo, O., Ebong, G.N., Ajagbe, S.A., A New Solution for Cyber Security in Big Data Using Machine Learning Approach, in: *Mobile Computing and Sustainable Informatics: Proceedings of ICMCSI 2023*, Springer Nature Singapore, Singapore, pp. 495–505, 2023.
34. Kozinets, R.V., *Netnography*, Wiley Online Library, USA, 2015.

31

Review of Research Methodology and IT for Business and Threat Management

Hitesh Rawat[1]*, Anjali Rawat[2], Sunday Adeola Ajagbe[3,4] and Yagyanath Rimal[5]

[1]*Department of Business Management and Economics, University of Extremadura, Badajoz, Spain*
[2]*Department of Computer and Communication Technology, University of Extremadura, Badajoz, Spain*
[3]*Department of Computer Science, University of Zululand, Kwadlangezwa, South Africa*
[4]*Department of Computer Engineering, Abiola Ajimobi Technical University, Ibadan, Nigeria*
[5]*Department of Computer Science, Pokhara University, Pokhara, Nepal*

Abstract

Technology is used to improve and streamline the research process, from data collection and analysis to reporting and distribution, through the combination of research methodology (REM) and IT. To enhance accuracy and efficiency, speed up data analysis, and simplify and automate activities, specialized software such as statistical analysis programs, data visualization tools, and collaboration platforms can be used. Furthermore, handling huge amounts of information and identifying intricate patterns may be facilitated by the use of cloud computing, big data analytics, and machine learning (ML). This enables Analysts to obtain fresh perspectives and draw more defensible conclusions. An organized REM for IT and threat management is represented by the work and focuses on methodical planning of data analysis and gathering strategies. Businesses may successfully discover possible risks, vulnerabilities, and security breaches by adhering to a well-defined research process. Businesses may proactively address security risks and safeguard their digital assets by putting in place a strong research and maintenance agreement (REM).

***Keywords*:** Research methodology, IT, business, threat management, data collection

**Corresponding author*: hrawat88@gmail.com

Rajesh Kumar Chakrawarti, Ranjana Sikarwar, Sanjaya Kumar Sarangi, Samson Arun Raj Albert Raj, Shweta Gupta, Krishnan Sakthidasan Sankaran and Romil Rawat (eds.) Natural Language Processing for Software Engineering, (475–486) © 2025 Scrivener Publishing LLC

Abbreviation Used

REM	Research Methodology
IT	Information Technology
ML	Machine Learning
QLR	Qualitative Research
QTR	Quantitative Research
MMR	Mixed Methods Research
NLR	Narrative Literature Review
SLR	Systematic Literature Review
MAR	Meta-Analysis Review
SLR	Scoping Literature Review
BM	Business Management

Table 31.1 REM framework [3].

Methods	Description
Qualitative research (QLR)	• collects and analyzes data regarding lived experiences, feelings, behaviors, and the meanings people ascribe to them using non-numeric techniques. • Analyst can get a deeper understanding of intricate ideas, interpersonal relationships, or cultural phenomena with the use of this kind of study.
Quantitative research (QTR)	• Measures variables and confirms preexisting theories or hypotheses using numerical data to produce numerical statistics. • Analyst typically use exams, databases, questionnaires, surveys, and organizational records to gather information from a significant sample of individuals. • comparisons and statistical analysis to examine the data.
Mixed methods research (MMR)	• integrates QTR and QLR. • To use QLR to investigate a scenario and create a conceptual framework, Analyst may employ QTR techniques to test the model experimentally. • Using a combination of closed-ended and open-ended surveys is an additional strategy where participants can create their own responses to certain questions and use pre-written answers for others.

31.1 Introduction

A strategy framework [1] called REM [2] describes the steps Analysts take to gather, evaluate, and analyze data. It serves as a roadmap for the whole investigation, guaranteeing the accuracy and authenticity of the results. Table 31.1 provides information about the REM framework.

- **Sampling designs in REM**

A key component of a REM is sampling, which is choosing a representative sample [4] of the population to research, drawing statistical conclusions about them, and estimating the features of the entire population from these conclusions. In REM, there are two different kinds of sampling designs: nonprobability and probability. The sampling design approaches are displayed in Table 31.2.

- **Data Gathering**

During research, data [6] are collected using various methods depending on the REM being followed and the research methods being undertaken.

Table 31.2 Sampling designs methods [5].

Method	Details
Probability sampling	• A sample is taken at random from a larger population. • Methodical sample members are selected on a regular basis and necessitate choosing a sample and sample size calculation beginning point that can be performed on a regular basis. • Clusters focus-population is separated using (age, sex, geography, demographic characteristics.)
Non-probability sampling	• The convenience approach uses this method to choose subjects who are most convenient for the Analyst to reach out based on factors like time of day, location, etc. • The purpose of the Analyst is to choose who gets to participate. • Analysts take into account the target audience's comprehension as well as the study's goal. • The snowball effect relates to the individuals who have previously been selected using their social networks to suggest new possible volunteers to the Analyst. • The number of participants in the study and their characteristics are determined by the Analyst during the design phase.

Table 31.3 Data gathering methods [7].

Method	Details
QLR	▪ 1- to-1 interviews: assist in comprehending a respondent's subjective viewpoint and experience on a certain subject or incident. ▪ Analyst examine pre-existing written resources, including reports, research papers, guidelines, official documents, and so on. ▪ For Learning the participants opinions on a particular issue, a moderator and a small sample of around 8–12 people generally participate in constructive conversations. ▪ QLR observation focusses on the five senses (sight, smell, touch, taste, and hearing) that are used by Analyst to gather data.
QTR	▪ Probability sampling most frequently used. ▪ Interviews - phone or in person. ▪ Observations -Analyzing people behavior in defined surroundings. ▪ Analyzing published papers. ▪ Includes questions and surveys- can be conducted via platforms (online or offline), based on the needs and size of the sample.

Both QLR and QTR have different data Gathering methods. Table 31.3 shows the Data Gathering Methods.

- **Data Analysis Methods**

For QLR and QTR, the data gathered [8, 11–13] using the different techniques must be analyzed to produce insightful findings. Between QTR and QLR, these data analysis techniques also vary. Table 31.4 provides information on data analysis techniques.

- **Literature Review**

The classifications of the literature reviews are displayed in Table 31.5.

- **Dataset**

A dataset in REM is a group of data that's used for modeling, analysis, or ML algorithm training. Although spreadsheets, tables, and databases are frequently used to organize datasets, they might simply include notes from QLR or photos or videos that were shot for a research project.

Table 31.4 Data analysis methods [8].

Methods	**Details**
QTR	• Incorporates a deductive approach to data analysis. • descriptive and inferential, uses statistical analysis software to examine numerical data. • descriptive analysis is used to characterize data types. • descriptive analysis methods: frequency measurements, central tendency measurements. • dispersion or variation measurements, • Position measurement (percentile and quartile rankings).
Inferential analysis	• It is employed to forecast the behavior of a broader group by analyzing data gathered from a smaller one. • The purpose of this analysis is to examine the connections between various variables. • To comprehend the link between two or more variables, use correlation. Examine the association between several variables using a cross-tabulation method. • Analyze the effects of independent factors on the dependent variable using regression analysis. • Frequency tables: To comprehend data frequency. • The purpose of variance analysis is to determine how much two or more variables differ from one another in an experiment.
QLR	• It uses an inductive approach to data analysis, in which theories are generated subsequent to data gathering. • Content analysis: This involves looking for certain words or concepts in texts to analyze recorded information from text and pictures. • Analyzing material from sources like surveys, field observations, and interviews is possible with narrative analysis. • Research issues are addressed by means of the narratives and viewpoints that individuals communicate. • Discourse analysis: This method examines social context, or the surroundings and way of life in which a conversation takes place, while analyzing interactions with individuals. • Grounded theory: formulates theories by gathering and evaluating facts in order to explain why a thing happened. • Thematic analysis is the process of finding significant themes or patterns in data and using them to solve problems.

Table 31.5 Literature review classifications [10].

Method	**Details**
Narrative Literature review (NLR)	• One type of research approach called a narrative review is a comprehensive summary of primary studies on a certain subject. • The review may be used to make methodical findings since it is grounded in the Analyst ‘s personal experience and current beliefs.
Systematic Literature review (SLR)	• Research is used to guide practice and decision-making based on evidence. • It entails looking for, assessing, and combining research findings. The methods used in systematic reviews might differ, and they are frequently study-specific.
Meta-analysis review (MAR)	• a statistically based systematic review that synthesizes the findings of separate investigations. • It can offer more accurate health care impact estimates than those obtained from the individual research included.
Scoping Literature Review (SLR)	• focuses on big research problems to gauge the amount of research that has been done. • It offers a first estimation of the possible volume and range of the body of research literature that is now accessible. • Key ideas, features of the literature, and research gaps may all be found with the use of scoping reviews.

Table 31.6 Datasets [6, 7].

Methods	**Details**
Numerical	• include numerals and are employed in QTR.
Text	• include documents, text messages, and postings.
Multimedia	• include music, video, and picture files
Time-series	• hold information gathered over time for trend and pattern analysis.
Spatial	• Include data that is spatially related, like GPS information.
Correlation	• include interrelated variables, meaning that altering one will affect all of the others.
Bivariate	• include a pair of variables that convey a connection between the information. • To illustrate how temperature variations relate to the time of day. • A dataset can comprise a temperature variable and a time variable.

Text, numbers, pictures, audio files, and even simple item descriptions can all be found in datasets. The details of the datasets are displayed in Table 31.6.

• REM for Business Management

REM is a fundamental framework [5, 6] for investigating and understanding various phenomena, with significant implications for business management. Its purpose lies in facilitating informed decision making, problem-solving, performance evaluation, and fostering innovation. Within the dynamic landscape of business management, REM shows about the evidence-based decision making, risk mitigation, strategic planning, and continuous improvement.

In the dynamic and increasingly complex landscape of business management, the importance of REM cannot be overstated. The Table 31.7 shows about the RM in BM.

• Threat management

Threat management [4–7] is an all-encompassing strategy that guards against cyber threats to an organization's networks, data, and systems

Table 31.7 RM in BM [8].

Method	Details
Evidence-Based Decision Making	• REM offers a paradigm for deriving actionable insights from large datasets in an era of abundant data. • empowering organizations to make well-informed decisions based on factual information rather than conjecture or gut feelings.
Risk Mitigation	• Businesses may reduce the risks associated with market instability, competitive challenges, and disruptive technologies by carrying out in-depth research and analysis. • By detecting possible risks and opportunities, REM facilitates proactive risk management, allowing companies to create backup plans and calculate reactions.
Strategic Planning	• Strategic planning is aided by REM's insightful analysis of consumer behavior, market trends, and competitive dynamics. • Businesses may achieve sustained development and competitive advantage by formulating strategic objectives. • outlining executable plans and allocating resources efficiently by utilizing strong research approaches.
Continuous Improvement	• REM encourages companies to embrace innovation, experimentation, and learning to promote a culture of continuous improvement. • Businesses are able to accommodate changing client wants and preferences by improving product and service offerings, streamlining operations, and refining strategy through iterative research cycles.

using a range of technologies and procedures. It entails locating, evaluating, ranking, and reacting to threats to lessen the effect of possible security events. Table 31.8 provides information on threat management systems.

- **REM tools**

Analysts [4] can get assistance from REM tools [6, 7] and software for activities like organizing references, doing data analysis, and creating data visualizations. Software for managing references: It facilitates the gathering,

Table 31.8 Threat management systems [7, 8].

Method	Details
Risk assessment	• To assist teams in understanding the state of their systems and creating a strategy to remedy vulnerabilities. • Threat intelligence is correlated with asset inventories and existing vulnerability profiles.
STRIDE model	• Identifying all ecosystem assets that may be attacked is part of the STRIDE attack [2, 3] strategy. • The rapid pace of digital change on a worldwide scale has hampered this process.
Asset identification	• Examine attack patterns and behaviors to find instances of them. • Expert systems, ML, and statistical techniques are frequently used in this discipline.
Attack detection	• a proactive strategy to reduce risk that aids security teams by charting. • the network's weak points so they can evaluate risk, find weaknesses, and take preventative action to safeguard important assets.
Attack path modeling	• a threat modeling reference model (REM) that aids in the identification and classification of possible security risks to software systems by developers and security experts. • This aids in their ability to identify possible security threats to a system and to prioritize security measures appropriately.
Threat analysis	• explains how possible enemies might make use of system flaws to further their objectives. • It recognizes risks and establishes a policy for mitigating them for a particular architecture, feature set, and configuration.

sharing, and organization of research. Table 31.9 provides information about the RM Tools.

Tools for statistical analysis and data visualization [1, 10] that assist Analyst in analyzing data aid in data visualization for academics. Details on the visualization tools are displayed in Table 31.10.

Table 31.9 RM tools [8, 9].

Tool	Details
Zotero:	• A free, browser-integrated application that assists Analyst in storing papers, publications, and research projects. • It features an organizing system for classifying, labeling, and labeling data.
Mendeley:	• It has the ability to import articles from other research applications and automatically create bibliographies.
EndNote:	• A single, comprehensive tool for organizing citations and references. • Analyst to format and construct bibliographies, track adjustments and changes, and exchange references with teams.

Table 31.10 Visualization tools [9, 10].

Tool	Details
SPSS	• It provides data analysis for descriptive and bivariate statistics, predictions, and numerical result forecasts. • Facilitates data processing, direct marketing, and graphing.
Tableau and Excel	• A data visualization tool

31.2 Conclusion

For business and threat management in IT, REM is essential. It entails gathering information, analyzing it, and drawing conclusions using methodical procedures. Businesses may recognize possible dangers, make well-informed choices, and create efficient risk-mitigation plans by utilizing a clearly defined REM. This method guarantees that IT expenditures are optimized for security and performance and are in line with business objectives. Data gathering and analysis in threat management research may be improved by implementing cutting-edge IT technologies. Analyst can more quickly spot patterns and trends by utilizing advanced techniques. This may result in improved threat mitigation and response plans. When combined with IT solutions, REM may offer insightful information and enhance decision-making procedures for companies dealing with a range of security issues.

References

1. Paul, J. and Criado, A.R., The art of writing literature review: What do we know and what do we need to know? *Int. Bus. Rev.*, *29*, 4, 101717, 2020.
2. Wohlin, C. and Runeson, P., Guiding the selection of research methodology in industry–academia collaboration in software engineering. *Inf. Softw. Technol.*, *140*, 106678, 2021.
3. Staron, M. and Staron, M., Action research as research methodology in software engineering, in: *Action Research in Software Engineering: Theory and Applications*, pp. 15–36, 2020.
4. Seniv, M.M., Kovtoniuk, A.M., Yakovyna, V.S., Tools for selecting a software development methodology taking into account project characteristics. *Radio Electron. Comput. Sci. Control*, 2, 1, 175–175, 2022.
5. Brachle, B.J., McElravy, L.J., Matkin, G.S., Hastings, L.J., Preparing leadership scholars in PhD programs: A review of research methodology training. *J. Leadersh. Educ.*, *20*, 3, 108–122, 2021.
6. Girard, N., Cardona, A., Fiorelli, C., Learning how to develop a research question throughout the PhD process: training challenges, objectives, and scaffolds drawn from doctoral programs for students and their supervisors, in: *Higher Education*, pp. 1–20, 2024.
7. Taekema, S. and van Klink, B., Progress in Legal Methodology–A Methodological Assessment of Six PhD Theses, in: *Law and Method*, pp. 1–24, 2023.
8. Mishra, A.K., Tyagi, A.K., Dananjayan, S., Rajavat, A., Rawat, H., Rawat, A., Revolutionizing Government Operations: The Impact of Artificial Intelligence in Public Administration, in: *Conversational Artificial Intelligence*, pp. 607–634, 2024.
9. Noonia, A., Beg, R., Patidar, A., Bawaskar, B., Sharma, S., Rawat, H., Chatbot vs Intelligent Virtual Assistance (IVA), in: *Conversational Artificial Intelligence*, pp. 655–673, 2024.
10. Rawat, R. and Rajavat, A., Perceptual Operating Systems for the Trade Associations of Cyber Criminals to Scrutinize Hazardous Content. *Int. J. Cyber Warf. Terror. (IJCWT)*, *14*, 1, 1–19, 2024, http://doi.org/10.4018/IJCWT.343314.
11. Nahar, S., Pithawa, D., Bhardwaj, V., Rawat, R., Rawat, A., Pachlasiya, K., Quantum technology for military applications, in: *Quantum Computing in Cybersecurity*, pp. 313–334, 2023.

12. Sikarwar, R., Shakya, H.K., Kumar, A., Rawat, A., Advanced Security Solutions for Conversational AI, in: *Conversational Artificial Intelligence*, pp. 287–301, 2024.
13. Pithawa, D., Nahar, S., Bhardwaj, V., Rawat, R., Dronawat, R., Rawat, A., Quantum Computing Technological Design Along with Its Dark Side, in: *Quantum Computing in Cybersecurity*, pp. 295–312, 2023.

About the Editors

Rajesh Kumar Chakrawarti, PhD is a dean and professor in the Department of Computer Science and Engineering, Sushila Devi Bansal College, Bansal Group of Institutions, India. He has over 20 years of professional experience in academia and industry. Additionally, has organized and attended over 200 seminars, workshops, and conferences and has published over 100 research papers and book chapters in nationally and internationally revered publications.

Ranjana Sikarwar is currently pursuing a PhD from Amity University, Gwalior. She competed her Bachelor of Engineering in 2006 and Master of Technology in Computer Science and Engineering in 2015. Her research interests include social network analysis, graph mining, machine learning, Internet of Things, and deep learning.

Sanjaya Kumar Sarangi, PhD is an adjunct professor and coordinator at Utkal University with over 23 years of experience in the academic, research, and industry sectors. He has a number of publications in journals and conferences, has authored many textbooks and book chapters, and has more than 30 national and international patents. He is an active member and life member of many associations, as well as an editor, technical program committee member, and reviewer in reputed journals and conferences. He has dedicated his career to taking care of Information and Communication Technology to enhance and optimize the information and worldwide research that can lead to improved student learning and teaching methods.

Samson Arun Raj Albert Raj, PhD is an assistant professor and placement coordinator in the Division of Computer Science and Engineering, School of Computer Science and Technology, Karunya Institute of Technology and Sciences, Tamil Nadu, India. His research is focused on smart city development using drone networks and energy grids with various applications, and his areas of expertise include wireless sensor networks, vehicular ad-hoc networks, and intelligent transportation systems.

Shweta Gupta, Assistant Professor, Computer Science and Engineering Department, Medicaps University, Indore (M.P.), India. I am Shweta Gupta work as an assistant professor of computer science and engineering at Medicaps University, with a focus on Natural Language Processing, Data mining, machine learning. I want to close the knowledge gap between theory and real-world applications in the tech sector through my love of research and teaching. My approach is centred on encouraging creativity and motivating students to strive for technological excellence.

Krishnan Sakthidasan Sankaran, PhD is a professor in the Department of Electronics and Communication Engineering at Hindustan Institute of Technology and Science, India. He has been a senior member of the Institute of Electrical and Electronics Engineers for the past ten years and has published more than 70 papers in referred journals and international conferences. He has also published three books to his credit. His research interests include image processing, wireless networks, cloud computing, and antenna design.

Romil Rawat has attended several research programs and received research grants from the United States, Germany, Italy, and the United Kingdom. He has chaired international conferences and hosted several research events, in addition to publishing several research patents. His research interests include cyber security, Internet of Things, dark web crime analysis and investigation techniques, and working towards tracing illicit anonymous contents of cyber terrorism and criminal activities.

Index

Also of Interest

From the same editors

Online Social Networks in Business Frameworks, Edited by Sudhir Kumar Rathi, Bright Keswani, Rakesh Kumar Saxena, Sumit Kumar Kapoor, Sangita Gupta, and Romil Rawat, ISBN: 9781394231096. This book presents a vital method for companies to connect with potential clients and consumers in the digital era of Online Social Networks (OSNs), utilizing the strength of well-known social networks and AI to achieve success through fostering brand supporters, generating leads, and enhancing customer interactions.

Quantum Computing in Cybersecurity, Edited by Romil Rawat, Rajesh Kumar Chakrawarti, Sanjaya Kumar Sarangi, Jaideep Patel, and Vivek Bhardwaj, ISBN: 9781394166336. This cutting-edge new volume provides a comprehensive exploration of emerging technologies and trends in quantum computing and how it is used in cybersecurity, covering everything from artificial intelligence to how quantum computing can be used to secure networks and prevent cyber crime.

ROBOTIC PROCESS AUTOMATION, Edited by Romil Rawat, Rajesh Kumar Chakrawarti, Sanjaya Kumar Sarangi, Rahul Choudhary, Anand Singh Gadwal, and Vivek Bhardwaj, ISBN: 9781394166183. Presenting the latest technologies and practices in this ever-changing field, this groundbreaking new volume covers the theoretical challenges and practical solutions for using robotics across a variety of industries, encompassing many disciplines, including mathematics, computer science, electrical engineering, information technology, mechatronics, electronics, bioengineering, and command and software engineering.

AUTONOMOUS VEHICLES VOLUME 1: Using Machine Intelligence, Edited by Romil Rawat, A. Mary Sowjanya, Syed Imran Patel, Varshali Jaiswal, Imran Khan, and Allam Balaram. ISBN: 9781119871958. Addressing the current challenges, approaches and applications relating to autonomous vehicles, this groundbreaking new volume presents the research and techniques in this growing area, using Internet of Things, Machine Learning, Deep Learning, and Artificial Intelligence.

AUTONOMOUS VEHICLES VOLUME 2: Smart Vehicles for Communication, Edited by Romil Rawat, Purvee Bhardwaj, Upinder Kaur, Shrikant Telang, Mukesh Chouhan, and K. Sakthidasan Sankaran, ISBN: 9781394152254. The companion to *Autonomous Vehicles Volume 1: Using Machine Intelligence*, this second volume in the two-volume set covers intelligent techniques utilized for designing, controlling and managing vehicular systems based on advanced algorithms of computing like machine learning, artificial Intelligence, data analytics, and Internet of Things with prediction approaches to avoid accidental damages, security threats, and theft.

Check out these other related titles from Scrivener Publishing

FACTORIES OF THE FUTURE: Technological Advances in the Manufacturing Industry, Edited by Chandan Deep Singh and Harleen Kaur, ISBN: 9781119864943. The book provides insight into various technologies adopted and to be adopted in the future by industries and measures the impact of these technologies on manufacturing performance and their sustainability.

AI AND IOT-BASED INTELLIGENT AUTOMATION IN ROBOTICS, Edited by Ashutosh Kumar Dubey, Abhishek Kumar, S. Rakesh Kumar, N. Gayathri, Prasenjit Das, ISBN: 9781119711209. The 24 chapters in this book provide a deep overview of robotics and the application of AI and IoT in robotics across several industries such as healthcare, defense, education, etc.

SMART GRIDS FOR SMART CITIES VOLUME 1, Edited by O.V. Gnana Swathika, K. Karthikeyan, and Sanjeevikumar Padmanaban, ISBN: 9781119872078. Written and edited by a team of experts in the field, this first volume in a two-volume set focuses on an interdisciplinary perspective on the financial, environmental, and other benefits of smart grid technologies and solutions for smart cities.

SMART GRIDS FOR SMART CITIES VOLUME 2: Real-Time Applications in Smart Cities, Edited by O.V. Gnana Swathika, K. Karthikeyan, and Sanjeevikumar Padmanaban, ISBN: 9781394215874. Written and edited by a team of experts in the field, this second volume in a two-volume set focuses on an interdisciplinary perspective on the financial, environmental, and other benefits of smart grid technologies and solutions for smart cities.

SMART GRIDS AND INTERNET OF THINGS, Edited by Sanjeevikumar Padmanaban, Jens Bo Holm-Nielsen, Rajesh Kumar Dhanaraj, Malathy Sathyamoorthy, and Balamurugan Balusamy, ISBN: 9781119812449. Written and edited by a team of international professionals, this groundbreaking new volume covers the latest technologies in automation, tracking, energy distribution and consumption of Internet of Things (IoT) devices with smart grids.

DESIGN AND DEVELOPMENT OF EFFICIENT ENERGY SYSTEMS, edited by Suman Lata Tripathi, Dushyant Kumar Singh, Sanjeevikumar Padmanaban, and P. Raja, ISBN 9781119761631. Covering the concepts and fundamentals of efficient energy systems, this volume, written and edited by a global team of experts, also goes into the practical applications that can be utilized across multiple industries, for both the engineer and the student.

INTELLIGENT RENEWABLE ENERGY SYSTEMS: Integrating Artificial Intelligence Techniques and Optimization Algorithms, edited by Neeraj Priyadarshi, Akash Kumar Bhoi, Sanjeevikumar Padmanaban, S. Balamurugan, and Jens Bo Holm-Nielsen, ISBN 9781119786276. This collection of papers on artificial intelligence and other methods for improving renewable energy systems, written by industry experts, is a reflection of the state of the art, a must-have for engineers, maintenance personnel, students, and anyone else wanting to stay abreast with current energy systems concepts and technology.

SMART CHARGING SOLUTIONS FOR HYBRID AND ELECTRIC VEHICLES, edited by Sulabh Sachan, Sanjeevikumar Padmanaban, and Sanchari Deb, ISBN 9781119768951. Written and edited by a team of experts in the field, this is the most comprehensive and up to date study of smart charging solutions for hybrid and electric vehicles for engineers, scientists, students, and other professionals.