2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS) | 979-8-3503-5968-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICKECS61492.2024.10616441

Enhancing Accuracy in Social Media Sentiment Analysis through Comparative Studies using Machine Learning Techniques

Kottala Sri Yogi[1]
Department of Operations
Symbiosis institute of Business Management Hyderabad
Symbiosis International University
Pune, Hyderabad, Telangana.
dr.sriyogi@outlook.com

Dankan Gowda V[2]
Department of Electronics and Communication Engineering,
BMS Institute of Technology and Management,
Bangalore, Karnataka.
dankan.v@bmsit.in

Sindhu D[3]
Assistant Professor, Department of Information Science and Engineering, B.G.S Institute of Technology-ACU, BG Nagara Pincode:571448, Karnataka, India.
sindhud@bgsit.ac.in

Hariprasad Soni[4]
Assistant Professor, Symbiosis Institute of Business Management, Hyderabad; Symbiosis International (Deemed University), Pune, India.
hr.soni@sibmhyd.edu.in

Saptarshi Mukherjee[5]
Mtech Bioinformatics, Makaut,Maulana Abul Kalam Azad University of Technology, Nadia West Bengal, Kolkata, India.

Madhu G.C[6]
Assistant professor, Department of ECE, Mohan Babu University (Erstwhile Sree Vidyanikethan Engineering College), Tirupati, India.
msnaidu417@gmail.com

Abstract: **- In the general scope of social media analytics, sentiment analysis is one of the most significant tools that can be employed to elucidate useful information from a vast amount of textual data. However, one of the primary problems that still persist with sentiment analysis is its accuracy because it is hard to understand precisely what a person meant on the vastness of the Web. In this research, these machine learning algorithms such as Naive Bayes (NB), Support Vector Machines and others are compared to determine its efficiency in enhancing the rate of accuracy in sentiment analysis across social media. Through data collection from various social media sources, which are preceded by stringent pre-processing techniques such as text normalization and feature extraction the performance of each model is evaluated. However, the findings present significant disparities in these models' accuracies, illustrating their best conditions of operation. This research helps to fill the gap in literature by offering a more subtle idea of strengths and weaknesses of every machine learning methodology when used for sentiment analysis applications, which is useful information on how researchers as well as practitioners can improve analytical accuracy within social media context.**

Keywords: Sentiment Analysis, Social Media Analytics, Machine Learning, Naive Bayes Classifier, Support Vector Machines (SVM), Text Pre-processing.

I. INTRODUCTION

Social media has seen tremendous growth in size and importance, developing into an integral component of the normal life. It provides an opportunity for people to share their ideas and beliefs, thus forming a database of information that can be used in research. With the help of Social Media Analytics (SMA) and Sentiment Analysis (SA), users' opinions can be quantified as well as qualitatively measured[1]. This section brings the notions of SMA and SA. Social media is referred to as web-based applications that are used for the easy creation, consumption and sharing of user generated content. Web 2.0 has changed the world of communication forever since it allowed people to connect via online platforms and social networks[2,3]. These platforms are used by businesses to promote personal opinions, products and services. Every day, the number of social media users increases and currently stands at 3.80 billion digital participants worldwide as of January 2023. Content posted on social media is a wide range of forms like text, videos, pictures and also music[4,5]. This type has created social media a potent instrument for getting and disseminating information throughout diverse areas of entertainment, business, science, politics as well as disaster management. One of the main reasons for social media's success is its cost-efficiency in broadcasting and receiving public messages, which has led to an increase in user engagement resulting into a massive accumulation of data that comes from text pictures videos audio geolocation information. Social media data can be divided into unstructured and structured types, including textual information on one hand, user relations as another type[6,7]. Social media usage has grown exponentially and enabled new fields of research, making it possible to explore social data in order to reveal what trends are current, public opinions or other types of informatio

979-8-3503-5968-8/24/$31.00 ©2024 IEEE

that would take surveys or focus groups on traditional basis[8,9]. Such analysis is comparable to qualitative research and hence social media becomes a central element in computational social science studies. In this case, quantitative computational methods such as statistics, machine learning with data mining and simulation modeling are used to study problems. Analytics is the foundation of any marketing plan, and in particular it becomes the most critical component when we talk about digital channel because of its vast ecosystem where platforms, advertising and promotion need to be meticulously measured[10].The figure.1. This provides a flowchart that depicts how the methods unfold over time and their interrelations in sentiment analysis with regards to social media.

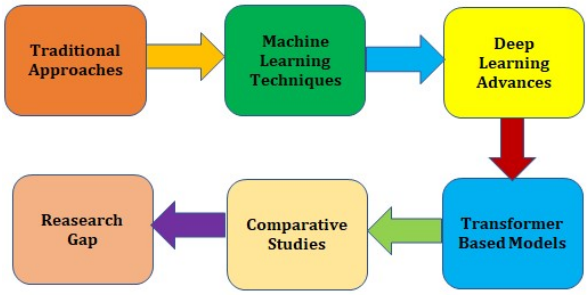

Fig.1: Techniques for Analyzing Sentiment on Social Media Platforms

In order to make the most out of marketing efforts, it is paramount that all actions are quantified and analyzed employing tools which improve on the quality of strategizing and planning[11,12]. SMA is a vital linkage in this ecosystem, which can be defined as the collection and analysis of data from social networks. Marketers usually use SMA for tracking discussions about products that are assessed online[13]. It provides companies with a wide variety of advantages, including increasing the perception of brands online, gaining market intelligence regarding new products to lead innovation while searching for opportunities from debates around particular areas such as stock and hashtags[14]. SMA finds its utility in several key marketing tasks: marketing effectiveness evaluation, defining product or service innovations process flow, tapping new customer segments; improving overall quality of services and innovation management to optimize R&D activities The emergence of SMA saw the rise in specialized data services tools and analytical platforms aimed at providing deeper market dynamics as well consumer behavior insight that helped marketers decide better.

II. LITERATURE SURVEY

Social media is a crucial medium in the modern society for communication, information sharing, news transmission and advertisement spreading. Its analytics are an important part of the process, helping to decode unstructured data and identify key application areas, trends as well as summarizing recent research findings[15,16]. Social media is an all-encompassing, inexpensive and pervasive communication channel that has reduced the world into a global village. It greatly impacts different aspects of life by demonstrating how people feel about products, services, topics and events as well as the fact that it has made most persons rely on such a source for news or information[17,18]. This means that the data collected via social media platforms is a vital tool in decision-making. The data from such platforms is collected by some entities to analyze it while others disseminate the information for benefits of a larger online community. Social media gives the users super powers by giving them a platform where they are free to share their opinions and emotions that affect organizations, public figures and individuals in equal measure[19,20]. Social media is a serious source of consumer feedback for organizations that use the information to redefine their products or services. Public figures work hard in managing their online profile to preserve it. Likewise, corporate bodies, political organizations and individuals also monitor social media activities to determine the public reaction on issues that concern them thus enhancing the importance of using social media in understanding peoples perception as well behavior.

For a comprehensive understanding of the opinions and sentiments expressed on SM sites by users, SA of authoritative content is required. SA uses text analysis and computational linguistics approaches to determine subjective knowledge from a broad array of marketing as well as customer service documents[21,22]. Its main objective is to distinguish the attitude of people towards different products or topics with a goal to uncover affective states, classify sentiment polarity (positive or negative) and obtain actionable information[23, 24]. These insights are critical for informed product related decision making.

In contrast, sentence-level SA is concerned with the extraction of sentiment polarity from sentences that are contained in a text [25]. In contrast, BNs unlike NB which assumes independence among entities use entity dependencies to determine the probability of distinct entities. DAG is used to show how variables are related. These are sophisticated topological structures developed by specialists and usually need cyclical improvement. BNs allow the derivation of unobservable factors from observed data, and their performance is demonstrated in SA tasks as well as IR. However, it is problematic because with more variables the number of possible DAG configurations increases dramatically [26]. This complexity also makes it challenging to devise an optimal topological structure for variables as well as the automation of learning DAG.

III. METHODOLOGY

This section is focused on introducing new methods that can enhance the effectiveness of Sentiment Analysis (SA) for text data from social media (SM), which categorizes these texts into positive, negative or neutral sentiments. Comprehending these emotions can help companies in the development of their marketing strategies, improvement of product offerings as well as in the enhancement of customer service.The figure.2. it provides a conceptual model that demonstrates three main components. This study outlines

the methodology that employed to analyze SM text data. The approach is systematic in terms of collecting, processing and analysis of the same. The process starts with the selection of a data source, which includes identifying appropriate social media sites and collecting textual content that reflects user sentiments towards certain topics or brands.

After the stage of data collection, this study uses a number of DM techniques in order to process and analyze text.

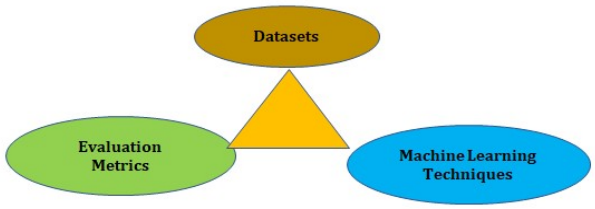

Fig. 2: Machine Learning Sentiment Analysis Strategies on Social Media Platforms

The last step involves using the trained and evaluated model to classification. Using this approach, organizations can benefit from SA and guide the public perception to make informed decisions based on customer behavior. This section not only presents an elaborate description of the methodology but also highlights that these techniques should be constantly refined and adapted to new social media trends and changing human language.

IV. DATASET

In the research, a Twitter sentiment dataset of 1.6 million tweets was used (Figure3.) This data set is obtained using the twitter API which shows people's content with diverse nature. Each tweet within the dataset has been annotated to indicate sentiment polarity: -1 for negative, 0 for neutral and +1 on positive sentiments. The variables that are incorporated in the dataset cover sentiment target (polarity of tweet), IDs, dates to which they were posted for each given tweet having a unique identifier whereby it is recorded when and by whom was collected using query if any. For pre-processing the dataset for analysis, several feature extraction techniques were used. Feature extractors could include tokenization methods, where the text is broken down into words or phrases; normalizing to standardize text by lowering case and removing punctuation marks; vectorizers such as bag-of-words or TFIDF that turn the texts into numerical vectors representing word significance in dataset.

V. IMPLEMENTATION

Categorization is a crucial process in the area NLP that enables to assign one or more pre-specified categories to a set of text documents according to their content. The application of this application is broad for instance in email filtering, categorization for improvement in search engine use and digital libraries classification. Regarding the present research, text categorization is used to change original text data into a form suitable for mining, emphasis being placed on features extracted from the texts that could distinguish the categories. The process of text categorization is critical and complicated. This step is designed to increase the relevance of words for tweets and association with particular categories. Doing so enables it to greatly contribute towards improving the accuracy and efficiency of categorization. The figure.4. shows the result of text classification. Although the figure is not presented here, one could imagine it representing how textual data from Twitter being initially unstructured and diverse gets organized into separate groups depending on its content. This visualization would aid in the comprehension of how effective were pre-processing steps and categorising algorithm to organize data meaningfully. Using such categorization, it is possible to analyze large sets of textual information effectively and extract insights, trends and patterns that would be tedious if done manually. This approach not only helps in the study of sentiment analysis on platforms such as Twitter but also increases text categorization use cases within NLP applications, which ultimately provides major advantages for information retrieval and content management.

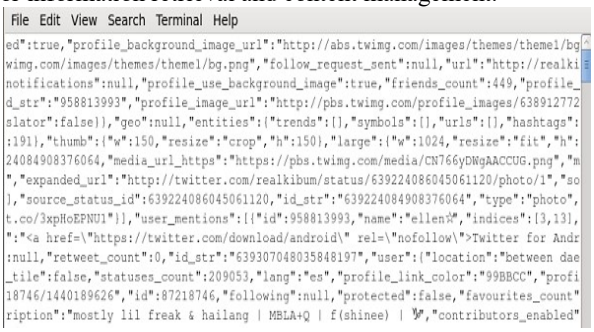

Fig.3: Tweets Database Snapshot

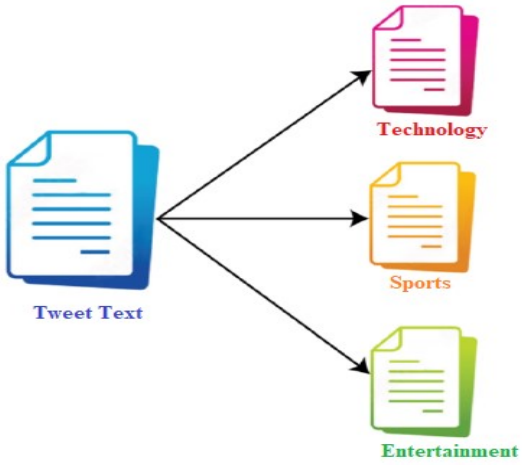

Fig.**4:** Text Categorization

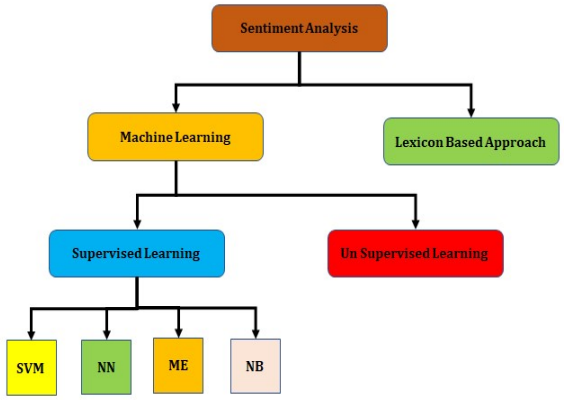

Fig.**5:** Overview of Techniques for Classifying Sentiment

Sentiment classification techniques are approaches (Figure.5.) that have been applied in NLP and text analysis for classifying the sentiment conveyed by a piece of content. These methods are designed to classify text into polarities of sentiment, for example positive and negative or sometimes more subtle categories like very positive or very negative. The first aim is to determine the emotional shade of a set of words in order to get some information about what opinions, attitudes and feelings were shared by the author or speaker. Figure.6. demonstrates how state-of-the art deep learning algorithms are used to analyze and interpret the sentiment contained in text data.

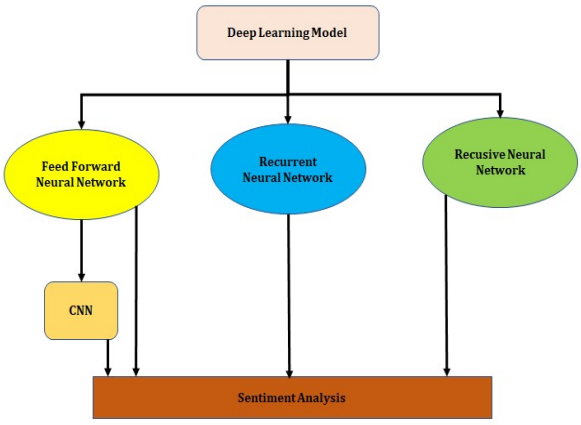

Fig.**6:** Implementation of Deep Learning in Sentiment Analysis

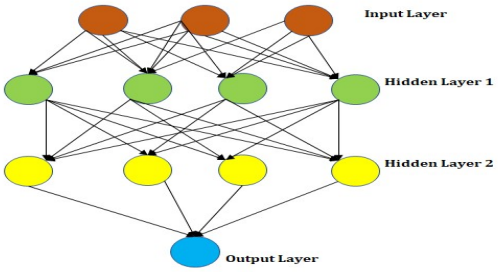

Fig.**7:** Schematic Representation of the MLP Network Structure

Figure.7. shows the intricate structure of a Multilayer Perceptron (MLP), an artificial neural network characterized by simplicity and efficiency in many machine learning tasks. The figure could also represent the flow of data across the network, emphasizing forward propagation to produce outputs from inputs and backpropagation for error correction on weights and biases, an important procedure required in training.

VI. RESULTS AND DISCUSSION

In the Results and Discussion section, we discuss our comparative study of machine learning techniques to improve accuracy in social media sentiment analysis. In this section, we evaluate the efficacy of different machine learning models such as Naive Bayes, Support Vector Machines Random Forest and Neural Networks on various social media datasets. Furthermore, we investigate heterogeneity in model performance across social media platforms and discuss the practical implications of our research for use as a tool in social media analytics. With this detailed analysis, we aim to present the findings that would help in understanding how effective machine learning techniques are at sentiment analysis and contribute towards improvements.

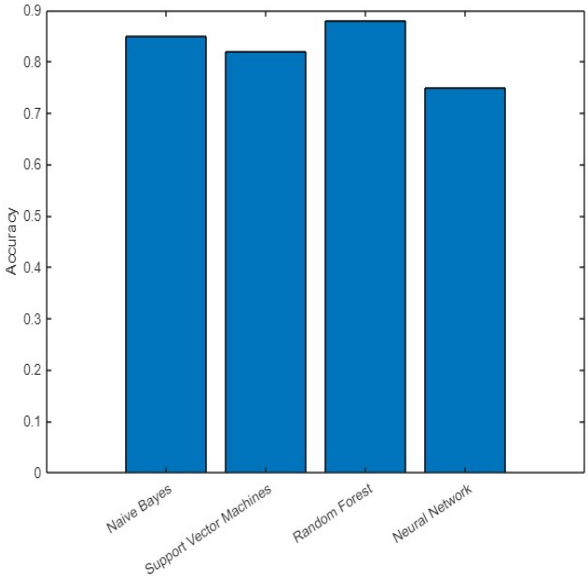

Fig.8: Comparative Performance of Machine Learning Techniques on Accuracy

Figure 8 depicts the comparative accuracy performance of different machine learning techniques employed in sentiment analysis on social media data. Each bar corresponds to the accuracy score obtained with a particular machine learning method, which makes it easy to compare their results. In Figure 9, precision-recall curves for two main machine learning methods in sentiment analysis: Naive Bayes and SVM are presented. These curves demonstrate the accuracy of each model in relation to recall variations, thus shedding light on their performance properties. Figure

10 shows the comparison of F1-scores attained by machine learning algorithms on Twitter, Facebook and Instagram. Each set of bars depicts the F1-score achieved by an approach on a particular platform, allowing comparison of model performance in various social media settings. The confusion matrices for Naïve Bayes and SVM are shown in Figure 11, which reflects the classification accuracy of these two methods used in sentiment analysis. In the matrix, each cell shows how many cases have been labeled by model providing an insight into what fraction of positive and negative sentiment cases were identified correctly with regards to this specific case.

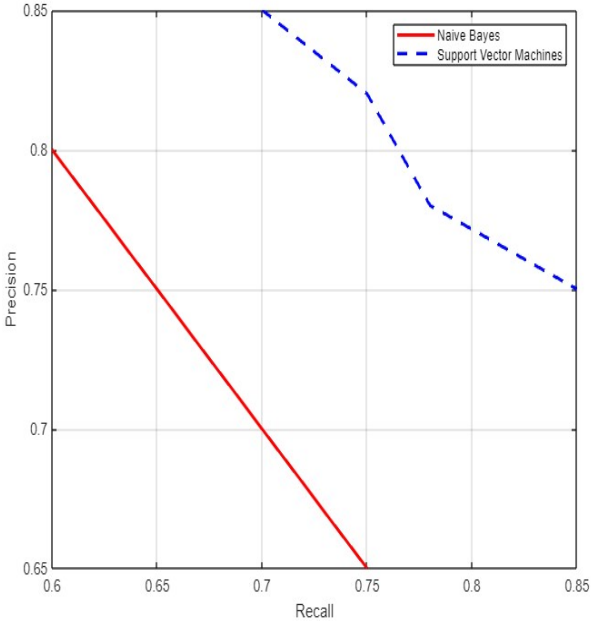

Fig.9: Precision-Recall Curve for Naive Bayes and Support Vector Machines

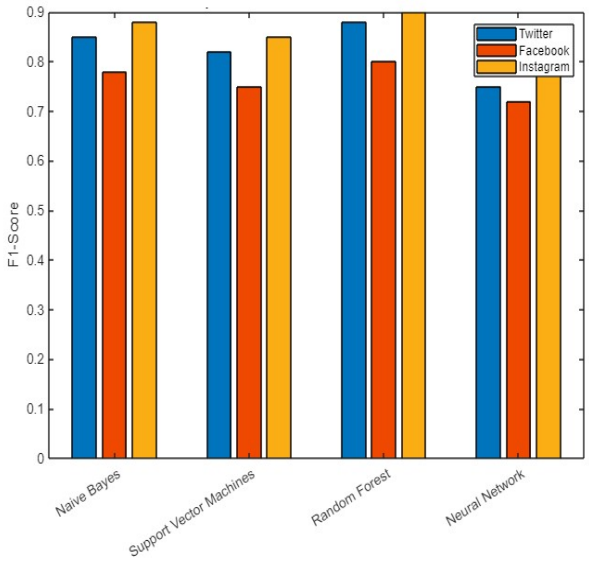

Fig.10: F1-Score Comparison across Different Social Media Platforms

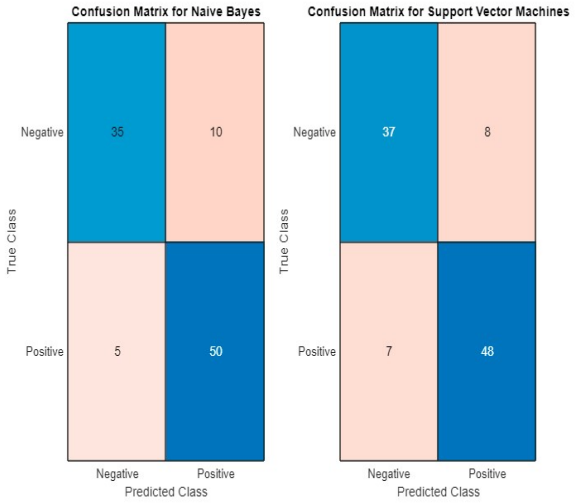

Fig..11: Confusion Matrix for Naive Bayes and Support Vector Machines

VII. CONCLUSION

In conclusion, this paper has performed a detailed comparative study of machine learning approaches to enhance the efficiency in social media sentiment analysis In various social media datasets, we have considered the effectiveness of some common techniques such as Naive Bayes, SVM, RF and NNs by conducting intensive data collection followed by strict assessment. As for the results we obtain, it is clear that there are significant variations in accuracy, precision and recall as well as F1-score for such approaches revealing their advantages and drawbacks of being applied as sentiment analysis tools. Moreover, our research has unearthed the pragmatic importance of these results with respect to social media analytics and given insights into the necessity to select appropriate machine learning models depending on dataset characteristics and platform peculiarities. This research also supports analytical method development in the socially dynamic area of analysis social media data by showing us how machine learning techniques function for sentimentality.

Going forward, future studies should build on the new methodologies and techniques to improve accuracy in sentiment analysis thereby contributing towards understanding online discourse and public opinion dynamics.

REFERENCES

[1] B. K. Bhavitha, and N. N. Chiplunkar, "Comparative study of machine learning techniques in sentimental analysis," 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 2017, pp. 216-221.

[2] K. Jain and S. Kaushal, "A Comparative Study of Machine Learning and Deep Learning Techniques for Sentiment Analysis," 2018 7th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2018, pp. 483-487.

[3] M. Rahman, "Optimizing Book Recommendations through Machine Learning: A Collaborative Filtering and Popularity-Based

Framework," 2023 4th IEEE Global Conference for Advancement in Technology (GCAT), Bangalore, India, 2023, pp. 1-8.
[4] M. T. H. K. Tusar, "A Comparative Study of Sentiment Analysis Using NLP and Different Machine Learning Techniques on US Airline Twitter Data," 2021 International Conference on Electronics, Communications and Information Technology (ICECIT), Khulna, Bangladesh, 2021, pp. 1-4, doi: 10.1109/ICECIT54077.2021.9641336.
[5] Y. Indulkar and A. Patil, "Comparative Study of Machine Learning Algorithms for Twitter Sentiment Analysis," 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 2021, pp. 295-299.
[6] S., Sarma, Parismita & Anand Kumar (2023) A novel approach of unsupervised feature selection using iterative shrinking and expansion algorithm, Journal of Interdisciplinary Mathematics, 26:3, 519-530.
[7] G. Kaur and A. Sharma, "Comparison of Different Machine Learning Algorithms for Sentiment Analysis," 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), Erode, India, 2022, pp. 141-147.
[8] L. Maada and Y. Farhaoui, "A comparative study of Sentiment Analysis Machine Learning Approaches," 2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), Meknes, Morocco, 2022, pp. 1-5.
[9] Hussain, Naziya (2023) A novel RF-SMOTE model to enhance the definite apprehensions for IoT security attacks, Journal of Discrete Mathematical Sciences and Cryptography, 26:3, 861-873.
[10] W. N. Chan and T. Thein, "A Comparative Study of Machine Learning Techniques for Real-time Multi-tier Sentiment Analysis," 2018 1st IEEE International Conference on Knowledge Innovation and Invention (ICKII), Jeju, Korea (South), 2018, pp. 90-93.
[11] N. S. L. S. Charitha, and J. S. Kiran, "Comparative Study of Algorithms for Sentiment Analysis on IMDB Movie Reviews," 2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2023, pp. 824-828.
[12] Selvakumar, & Chaturvedi, Abhay (2023) Securing networked image transmission using public-key cryptography and identity authentication, Journal of Discrete Mathematical Sciences and Cryptography, 26:3, 779-791.
[13] K. Dhola and M. Saradva, "A Comparative Evaluation of Traditional Machine Learning and Deep Learning Classification Techniques for Sentiment Analysis," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2021, pp. 932-936.
[14] Pullela SVVSR Kumar and B. Ashreetha (2023), Performance Analysis of Energy Efficiency and Security Solutions of Internet of Things Protocols. IJEER 11(2), 442-450.
[15] Poornima and K. S. Priya, "A Comparative Sentiment Analysis Of Sentence Embedding Using Machine Learning Techniques," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020, pp. 493-496.
[16] Bhasin and S. Das, "Twitter sentiment analysis using Machine Learning and Hadoop: A comparative study," 2021 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC), Jalandhar, India, 2021, pp. 267-272.
[17] K. S and M. R. Arun, "Priority Queueing Model-Based IoT Middleware for Load Balancing," 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS), *2022, pp. 425-430.*
[18] Takrim UL and Ramesha M (2022), Human Emotion Recognition using Deep Learning with Special Emphasis on Infant's Face. IJEER 10(4), 1176-1183. DOI: 10.37391/IJEER.100466.
[19] A. M. Reddy, "Improved Secure Communication Mechanism for IoT Platform Devices," 2022 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON), Bangalore, India, 2022, pp. 1-6.
[20] S. Reddy P, P. S. Patwal, "Data Analytics and Cloud-Based Platform for Internet of Things Applications in Smart Cities," 2022 International Conference on Industry 4.0 Technology (I4Tech), 2022, pp. 1-6.
[21] N. S. Reddy and B. Ashreetha, "Technologies for Comprehensive Information Security in the IoT," 2023 International Conference for Advancement in Technology (ICONAT), Goa, India, 2023, pp. 1-5.
[22] P. Pavankumar, N. K. Darwante, "Performance Monitoring and Dynamic Scaling Algorithm for Queue Based Internet of Things," 2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), *2022, pp. 1-7.*
[23] A. Singla, N. Sharma, "IoT Group Key Management using Incremental Gaussian Mixture Model," 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC), 2022, pp. 469-474.
[24] J. Singh and G. Sharma, "Sentiment Analysis Study of Human Thoughts using Machine Learning Techniques," 2023 International Conference on Disruptive Technologies (ICDT), Greater Noida, India, 2023, pp. 776-785.
[25] LS Chen, CH Liu and HJ Chiu, "A neural network based approach for sentiment classification in the blogosphere", journal of informetrics, vol. 5, no. 2, pp. 313-322, 2011.
[26] M. Thelwall *and* G. Paltoglou, "Sentiment Strength Detection for the Social Web", Journal of the American Society for Information Science and Technology, vol. 63, no. 1, pp. 163-173, January 2012.