**Detection Process of Suspicious Activities on Social Media Using Data Mining and Machine Learning**

CSE 543, Information Assurance and Security, Fall 2023, Group 1-5

A Project Report

Presented to

Professor Stephen Yau,

School of Computing and Augmented Intelligence



By

Krupaben Kothadia (GL), [kkothadi@asu.edu](mailto:kkothadi@asu.edu)

Gautham Vijayaraj (DL), [gvijaya6@asu.edu](mailto:gvijaya6@asu.edu)

Avani Mundra, [amudra@asu.edu](mailto:amudra@asu.edu)

Justin Young, [jtyoun15@asu.edu](mailto:jtyoun15@asu.edu)

Anuranjan Dubey, [adubey37@asu.edu](mailto:adubey37@asu.edu)

Rahul Nayak, [rrnayak@asu.edu](mailto:rrnayak@asu.edu)

Sangeeth Santhosh, [ssantho9@asu.edu](mailto:ssantho9@asu.edu)

Yeshwanth Reddy Chennur, [ychennur@asu.edu](mailto:ychennur@asu.edu)

**TABLE OF CONTENTS**

1. [**INTRODUCTION**](#_r2ojbcadpzat)**6**
   1. [Motivation and Background](#_ffabc9ytxb4d) 6
   2. [Goals and scope of study of this project](#_qda2n4nf74au) 7
2. [**SUMMARY AND ACCOMPLISHMENTS OF THE PROJECT**](#_2z0y3lw1bi2u) **9**
   1. [Data Exploration](#_kyeibjmev3mn) 9
   2. [Data Preprocessing](#_7lfc6q4k3bsh) 9
   3. [Secure Data Collection and Privacy Preservation](#_fudju61agcd9) 9
   4. [Data Mining](#_p3n3t18t5dfo) 9
   5. [Machine Learning Techniques](#_zad3gflx5kt5) 10
   6. [Model Security and Evaluation](#_n4lh58b5j5pq) 10
   7. [Accurate and Efficient Detection](#_hznwl9qk0fjc) 10
   8. [Adaptive Learning and Insightful Models](#_ka263qs8agww) 10
3. [**ACCOMPLISHMENTS OF EACH GROUP MEMBER**](#_dv0tvostzxwv) **11**
   1. [Krupaben Kothadia](#_mqo6yhcm0jzt) 11
   2. [Gautham Vijayaraj](#_npbgyhburnmj) 11
   3. [Avani Mundra](#_z839hk872q98) 12
   4. [Justin Young](#_ojrzy4j2e871) 13
   5. [Anuranjan Dubey](#_ogg5z7kmle1e) 14
   6. [Rahul Nayak](#_tefrr1a9ncoi) 14
   7. [Sangeeth Santhosh](#_l6hxhn16lvs5) 15
   8. [Yeshwanth Reddy Chennur](#_w8w4aa6s8sa7) 16
4. [**DETAILED RESULTS OF ALL INDIVIDUAL GROUP MEMBERS**](#_nfvwwgc2l5s6) **16**
   1. [Krupaben Kothadia](#_4q3pk3z8tcjw) 16
   2. [Gautham Vijayaraj](#_au73vcigfewj) 22
   3. [Avani Mundra](#_n0dyg4aa1u98) 27
   4. [Justin Young](#_or8x3rl6opd2) 32
   5. [Anuranjan Dubey](#_70597634dv3u) 37
   6. [Rahul Nayak](#_cfmtvb4vq63u) 42
   7. [Sangeeth Santhosh](#_h5l9sb3hvnev) 49
   8. [Yeshwanth Reddy Chennur](#_7lvr1ae31wlo) 54
5. [**CONCLUSIONS AND RECOMMENDATIONS**](#_kuxor3gcn66k) **41**
6. [**REFERENCES**](#_rjjmr38i4jor) **69**

**LIST OF FIGURES**

| Index No | Figure Number | Description | Page Number |
| --- | --- | --- | --- |
| 1 | Figure 4.1.1 | Mean Accuracy of Classifiers | 18 |
| 2 | Figure 4.1.2 | Figure shows the tags that are used to create POS tags. | 19 |
| 3 | Figure 4.2.1 | Different approaches to privacy-preserving data mining and the trade-off between efficiency, security and accuracy | 24-25 |
| 4 | Figure 4.3.1 | Extremist content detection | 31 |
| 5 | Figure 4.3.2 | Text mining process | 32 |
| 6 | Figure. 4.4.1 | System Design Flow | 32 |
| 7 | Figure. 4.4.2 | Methodology | 33 |
| 8 | Figure. 4.4.4 | The proposed model. | 34 |
| 9 | Figure. 4.4.5 | Model creation using UNSW dataset. | 35 |
| 10 | Figure. 4.4.6 | Recall, precision and F1 of SPC model compared to complex decision and ensemble boosted tree. | 35 |
| 11 | Figure 4.5.1 | Bidirectional LSTM architecture for named entity recognition. | 37 |
| 12 | Figure 4.5.2 | Graphical representation of RDF for tweets by SVCE. | 39 |
| 13 | Figure 4.5.3 | A framework for monitoring and analyzing tweets related to cyber attacks. | 39 |
| 14 | Figure 4.5.4 | Architecture of the proposed model. | 42 |
| 15 | Figure 4.6.1 | Types of Suspicious activities on social media | 43 |
| 16 | Figure 4.6.2 | Flowchart depicting types of features extracted from twitter | 45 |
| 17 | Figure 4.6.3 | Table depicting results | 48 |
| 18 | Figure 4.7.1 | Naive Bayes and Random Forest Algorithm Classification Results | 50 |
| 19 | Figure 4.7.2 | Comparison of Results of different algorithms | 51 |
| 20 | Figure 4.7.3 | Implementation Process of Proposed Approach | 54 |
| 21 | Figure 4.8.1 | Illustration of defensive techniques of machine learning | 55 |
| 22 | Figure 4.8.2 | A trusted social network using hypothetical decision making model | 56 |
| 23 | Figure 4.8.3 | System model of the proposed SDN-based anomaly detection framework | 57-58 |
| 24 | Figure 4.8.4 | Rumor detection flowchart | 59 |
| 25 | Figure 4.8.5 | Classification Flow Chart | 61 |

**LIST OF TABLES**

| Index No | Table Number | Description | Page Number |
| --- | --- | --- | --- |
| 1 | Table 4.2.1 | A summary of data anonymization approaches | 23-24 |
| 2 | Table 4.3.1 | Comparison of embedding layers | 28 |
| 3 | Table 4.3.2 | Comparison of classifier | 28 |
| 4 | Table. 4.4.1 | Result Accuracy | 33 |
| 5 | Table. 4.4.2 | Result Summary | 34 |
| 6 | Table 4.4.3 | Performance comparison of seven models-1 | 36 |
| 7 | Table 4.4.4 | Performance comparison of seven models-2 | 36 |
| 8 | Table 4.5.1 | Results for supervised machine learning algorithms. | 42 |
| 9 | Table 4.7.1 | Word Frequency for Opinion Spam (VS) Legitimate User | 52 |
| 10 | Table 4.7.2 | Comparison of Performance of Different Algorithms | 53 |

##### **INTRODUCTION**

###### **Motivation and Background**

Did you know that without proper data security, your social media accounts are vulnerable to cloning and other threats? There are many risks with social media and security, such as hacking, fake bots, account cloning, and more. The need to solve important problems quickly in data security, data mining, different social media platforms, and machine learning is what drove this project. The seven study areas we chose were based on our dedication to meeting the project's main goal. These include Data Exploration, Safe Data Collection, Data Preprocessing/NLP, Data Mining, Machine Learning Models, Model Security Assurance, and Model Evaluation.

The alarming rise in phishing scams, which has been reported on several times by The New York Times, shows how determined people are to fight data security risks. Scams like these are very dangerous for people, businesses, and society as a whole. We want to find suspicious content on social media sites by digging deeper into Data Exploration. This will help make the internet a better place.

The Secure Data Collection techniques make our project safer by building on a lot of study on existing methods. We focus on Data Preprocessing, Natural Language Processing (NLP), and Data Mining. This allows us to improve system robustness through careful paper exploration, creating a safer and more reliable environment.

Along with that, our talks about machine learning models, their safety features, and strict testing methods help to build a stronger base for reliable machine learning models, which is important for dealing with the changing types of data risks.

In today's digital world, the relationship between social media and safety is getting more complicated and important. As the world gets smaller and smaller, the risks to data and personal information are always there. An enormous rise in the number of online platforms and the widespread use of social media have made it possible for many threats to reach both people and businesses. Phishing scams try to trick people into giving up private information, and account cloning is becoming a bigger problem. Strong data security steps have never been more important. In this way, the project becomes important in both the academic and real worlds, making the internet a better place.

###### **Goals and scope of study of this project**

The goals of the project are as follows:

* **Build Sturdy Machine Learning Models:** The main objective is to build trustworthy machine learning models that can recognize a range of dubious activities on social media sites, such as hate speech, cyberbullying, fraudulent accounts, and false information, among other things. It could be necessary to design several models for this.
* **Attain Excellent Accuracy and Precision:** The project's goal is to deliver exceptional research outcomes by using machine learning models that reliably detect suspicious activity on social media with high degrees of accuracy, precision, recall, and F1-score.
* **Effective Processing of Data:** The focus of the study will be on finding effective data processing methods that can handle massive data quantities without sacrificing efficiency.
* **Perform Comprehensive Analysis:** To pinpoint particular traits and indicators that point to questionable activity, a thorough analysis will be carried out. This will provide light on the rationale behind the classification of questionable activity.
* **Facilitate Adaptive Learning:** The goal of the project is to create a system that can change and grow in order to recognize new dangers and questionable activity on social media.
* **Provide Explainable Predictions:** The machine learning models will not only produce precise forecasts but will also provide justifications for the suspicious label placed on a given activity.
* **Ensure Privacy-Preserving Techniques:** The project's main goal is to develop a system that actively detects suspicious activity on social media, protects user data privacy, and upholds ethical norms.
* **Record and Report Project Findings:** A thorough report describing the project's methods, algorithms, stages of execution, and results will be written.
* **Establish the Foundation for Additional Research:** Future studies in the area of social media behavior detection will be built upon the models and study results that have been produced.

Our project aims to research in topics related to social media suspicious activity detection using machine learning by leveraging a multifaceted approach encompassing the following key areas:

* **Data Exploration:** Here we will explore diverse social media suspicious activities, their manifestations, and potential implications, informed by relevant research literature.
* **Secure Data Collection:** We will study and employ secure data collection techniques that uphold data confidentiality and integrity while dealing with real-world social media data.
* **Data Preprocessing and Feature Engineering:** By leveraging various techniques, including NLP methodologies, we will research on different data preprocessing and feature engineering techniques that have been used for detecting suspicious activities.
* **Data Mining Techniques for Multimodal Data:** To effectively process and analyze multimodal social media data (text, images, videos, etc.), we will explore different data mining techniques that are used to detect suspicious activities
* **Machine Learning Models:** By assessing the adaptability of existing machine learning algorithms to detect suspicious activities on different social media platforms with minimal modifications, we will investigate whether these models can be extended beyond their initial implementations to ensure their applicability across diverse social media networks.
* **Model Security Assurance:** Recognizing the need for secure machine learning detection models, we will engage in discussions and research surrounding techniques to fortify these models against tampering or unauthorized modifications by potential hackers.
* **Model Evaluation:** To uphold the trust and reputation of honest users, our research will be dedicated to identifying the most suitable approach for classifying social media accounts into suspicious and non-suspicious categories without any false grouping, thereby ensuring the integrity and accuracy of our project.

This comprehensive project scope embodies a holistic strategy for detecting suspicious activity on social media, spanning from initial data exploration to model evaluation, with a keen focus on ethics, user privacy, and the adaptability of machine learning solutions across diverse social media platforms. By addressing these critical areas, our project aims to contribute to the enhancement of online security and the protection of honest users in the digital realm.

##### **SUMMARY AND ACCOMPLISHMENTS OF THE PROJECT**

###### **Data Exploration**

Data exploration research domain was focused on exploring types of social media data available on various platforms. In this domain we have explored spam or fake, and criminal datasets available on social media platforms like facebook, twitter, instagram and many more. The data exploration research has also been performed on Arabic language dataset. The research provides an in-depth analysis of applying machine learning to discern spammers on the Weibo platform and sheds light on the most indicative features for modeling this classification problem.We have discussed an approach to identify and curb Sybil attacks on social networks. Sybil attacks, a kind of malicious activity in online communities, can be detected and mitigated through the machine learning techniques described. The research demonstrates how applying this analytical framework can counteract a prevalent threat, strengthening trust in these spaces.

###### **Data Preprocessing**

Data preprocessing research domain also discusses the various NLP techniques acquired after preprocessing data. Here, we have discussed various data preprocessing and NLP techniques like pre-classification of text, cleaning, standardization, Tokenization, vectorizing, sequence analysis translating, prediction, a language model that classifies tweets as spam or not based on text-URL divergence concept, stemming and diacritics in the preprocessing of the data, LSB - steganography, sentiment analysis and VADER sentiment analysis.

###### **Secure Data Collection and Privacy Preservation**

Secure Data Collection and Privacy Preservation techniques identified in this research domain are distortion, encryption, and anonymity, Secure multiparty computation allows limited data access, tweet polarity values, privacy-preserving data mining (PPDM) techniques based on their data modification approaches, detecting distinct behavioral patterns through outlier detection paired with visual explanation, assignment of risk scores to users, and a privacy-preserving approach 3LP+ are discussed.

###### **Data Mining**

In this research domain we have discussed various data mining methods which have enhanced the performance of suspicious activities detection are: the similarity metric, a comparative analysis of word embedding models and classifiers, Word2Vec with bi-functionality, incorporation of emotional, sentiment, domain-specific lexicons and TFIDF embedding.

###### **Machine Learning Techniques**

The machine learning technique is one of the crucial parts of our research which discusses various ML techniques to accurately identify suspicious activities on social media. ML techniques discussed here are: classification methods utilizing deep neural networks (CNN and BiLSTM), Random Forest Regression, Logistic Regression, Support Vector Machines, and various deep learning algorithms.

###### **Model Security and Evaluation**

In this area of domain, research papers collectively address key aspects of enhancing online security and trust. They introduce innovative methods to foster trust in social networks, bolstering content filtering and prioritization. Mobile device security is emphasized, with various machine learning methods achieving high accuracies. Additionally, the efficiency of machine learning-based intrusion detection systems in enhancing social media security is demonstrated, particularly in detecting specific attacks. Furthermore, a novel "transfer learning" approach is introduced for cloud security, achieving high accuracy in classifying attacks and identifying unknown threats, underlining the importance of cloud security and context-based anomaly detection.

###### **Accurate and Efficient Detection**

Studies in this area collectively address various aspects of social media security and content filtering. They explore topics such as spam detection, efficient detection methods, safeguarding user integrity, and phishing analysis on platforms like Twitter. Different algorithms and approaches are employed to ensure accurate detection and classification of content, emphasizing the importance of maintaining the security and integrity of legitimate users while mitigating threats and suspicious activities in the dynamic realm of social media.

###### **Adaptive Learning and Insightful Models**

Research papers in this domain address security threats and solutions using machine learning models. They emphasize the importance of automated real-time detection through adaptive learning. Some papers focus on identifying insightful models, suggesting the consideration of various metrics. They explore the Decision Tree model for intrusion detection, demonstrating its superior performance in terms of accuracy and training time. The SPC model is used for enhanced security in cloud computing, outperforming conventional models. Additional research underscores the need for algorithms and models to mitigate suspicious activities in social networks, while proposing a deep learning-based approach for improved anomaly detection in multimedia applications.

##### **ACCOMPLISHMENTS OF EACH GROUP MEMBER**

###### **Krupaben Kothadia**

* In order to build a reliable data mining and machine learning pipeline, and to extract important insights from the data, we concentrate on data preprocessing and natural language processing (NLP). To accomplish this task of determining the appropriate strategies, seven research publications were referred.
* In [10], to preserve privacy of users, the authors analyzed detrimental tweets using cutting-edge NLP algorithms. They gathered data from Twitter, evaluated and categorized messages, and created a language model that used "divergence point" to define spam.
* In [12], because Arabic has a distinct structure, the study presented difficulties while analyzing tweets in the language. They gathered and handled stemming and diacritics in the preprocessing of the data. Tweets were manually categorized into two groups: suspicious and not suspicious.
* In [13], To identify dangers connected to terrorism in social networks, this study combines NLP and LSB - steganography technique approaches. It includes sentiment analysis, to find suspicious behavior patterns.
* In [11], sentiment analysis for short Twitter and Facebook texts is discussed. Data preprocessing included VADER sentiment analysis, standardization, and cleaning.
* In [14], it denotes that by examining the subjects of messages and communication patterns, it is possible to identify suspicious social network users. Tokenization, prediction, and sequence analysis—including the identification of "bad words"—are all accomplished. It suggests ways to block people that pose a threat.
* In [43], the authors build similarity scores for tweets by cleaning, normalizing, translating, and vectorizing them. The approach was evaluated on two protest scenarios, using different measures.
* In [42], to identify cybercrimes in social media, the writers go over the cleaning, normalization, translation, and vectorization processes. The analysis of Twitter data using the model is presented in the study along with its evaluation.

###### **Gautham Vijayaraj**

* It is important to find the right balance between harnessing efficient data mining techniques and respecting the user's privacy to maintain data confidentiality. 7 research papers were referenced to determine such data mining techniques.
* In [5], different methods of maintaining privacy while performing data mining were explored. These privacy preserving data mining algorithms addressed were data distribution, data distortion, association rule mining and data anonymity.
* In [6], an elaborate study on existing challenges and unresolved issues in privacy-preserving data mining (PPDM) was conducted. The existing PPDM techniques are intensively reviewed and classified based upon their methods that use data modification approaches.
* In [7], the main goal was to find how technology from the security community can change data mining for the better while still maintaining privacy. Approaches like data perturbation, randomization and secure multiparty computation were discussed.
* In [8], an approach for anomaly detection in Online Social Networks (OSNs) using data mining techniques was discussed. Anomalous users in social networks were detected based on their distinct behavioral patterns using outlier detection along with a visual explanation.
* In [9], the paper focused on cultivating an approach to analyze the sentiments of users using data mining classifiers. The sentiment analysis was performed using the k-nearest neighbor classifier which achieved an accuracy of 99.6456%.
* In [40], the growing concern of privacy in online OSNs where users often share personal information was addressed. The authors introduce a privacy-preserving technique called 3LP+ to safeguard multiple sensitive attributes that users may have.
* In [44], the primary objective is to use intelligent data mining techniques to extract valuable insights from user-generated content on social media platforms, ultimately enhancing healthcare outcomes. This can be applied to the overall goal of our project.

###### **Avani Mundra**

* Textual data is highly unstructured and complex and thus efficient data mining techniques are necessary to derive meaningful information from the multimodal data and apply efficient machine learning algorithms to identify suspicious content in social media.
* In [15], the authors create a framework using Word2Vec embedding layer and Multilayer Perceptron neural network model with 0.84 F1 score to detect abusive content in social media by analyzing Twitter dataset. Data mining techniques have been implemented to preprocess text for this framework.
* In [16], the study involved implementing a text mining technique using the Normalized Compression Distance function and exploiting the usage of ‘hashtags’ in Twitter to detect malicious content in social media.
* In [17], several word embeddings and deep learning networks have been compared and to identify the best framework with highest F1 score which comprises Random Forest Algorithm with TFIDF embedding to detect cyberbullying, an escalating issue with the ever-growing social media base.
* In [38], the authors present a comprehensive big data mining framework and employ sentiment analysis and NLP algorithms after processing text through data mining methods to specifically detect extremist content in large social networks.
* In [39], diverse user behavior patterns have been analyzed using the technique of process mining and recurrent neural networks to predict abnormal user behavior in social media platforms.

###### **Justin Young**

* There are many different ways in which your data security can be compromised on social media. Machine learning-based methods are powerful tools that can be used to ensure security on social media. Five research papers were referred to demonstrate the performance of machine learning models in various different environments.
* In [28], it covers a weakness of machine learning approaches to more sophisticated malware approaches in mobile security environments. This paper proposes a process to determine which machine learning algorithms are most efficient in a behavior-based malware detection system. This process is performed on the Naive Bayes, K-Nearest Neighbor (KNN) and Decision Tree algorithms.
* In [26], this paper performs a comparative analysis of machine learning models in computer network intrusion detection. The models analyzed in this research are XG-Boost, Decision Tree, Random Forest, KNN, Multi-Layer Perceptron (MLP), and Quadratic Discriminant Analysis (QDA).
* In [27], it proposes a new network intrusion detection model called the Secure Packet Classifier (SPC) for cloud intrusion detection. This model aims to detect and classify cloud anomalies using collaborative filtering between two machine learning algorithms.
* In [37], it proposes another technique, transfer learning, which can be used in the cloud to detect and classify malicious activity. In testing this method paired with a deep learning model, it proved to be effective in classifying known attack types and detecting unknown attacks in the cloud.
* In [29], It provides research on SQL injection and XSS cross site scripting, and a comparative analysis of machine learning models against these types of attacks. In the SQL-based experiment, the Adaboost model proved most accurate. In the XSS experiment, the SVM model had the highest accuracy and detection rate.

###### **Anuranjan Dubey**

* Deep neural networks, such as CNN and BiLSTM, are implemented in [22] to classify security-related content on Twitter, thereby streamlining the collection of threat intelligence and providing security analysts with practical value.
* The SVCE is utilized in [19] to extract entities associated with cybersecurity vulnerabilities from Twitter data. The CyberTwitter system exhibits promise in delivering significant threat alerts and generates customized notifications.
* It compares several algorithms for classifying security content on Twitter in [20], where Expectation Regularization is identified as the most promising. The research emphasizes Twitter's utility as a repository of information regarding security events.
* The method described in [21] employs feature extraction and the Random Forest machine learning technique to accurately detect phishing URLs on Twitter using authentic datasets. This guarantees prompt notifications to users.
* A feature-based method for identifying fraudulent Twitter accounts is described in reference [36]. This approach makes use of machine learning algorithms and 24 unique features. Fake account detection is improved through the examination of diverse attributes of accounts, tweet content, ownership particulars, and embedded URLs; Random Forest demonstrates the most exemplary performance in this regard.
* In addition, the not so important papers contained additional machine learning algorithms, such as those pertaining to the prediction of social media crime and the detection of Instagram threats utilizing Inception v3, as well as time series-based machine learning techniques. These papers provide a comprehensive understanding of field topics.

###### **Rahul Nayak**

###### In order to dive deep into detecting suspicious activities , it is important that we study the different types of suspicious activities that happen on social media.Thus my research domain delves deep into studying different types of suspicious activities

* Paper [4] explores different types of suspicious activities on social media, explains their characteristics, and discusses their impact on social media.
* Paper [2] proposes a machine learning framework for spam detection on Twitter, highlighting the importance of feature selection and discussing the challenges of detecting evolving spam tactics.
* Paper [3] addresses the identification and mitigation of malicious social media profiles using machine learning models to achieve high accuracy in distinguishing between genuine and malicious accounts.
* Paper [1] focuses on detecting spam accounts on Sina Weibo. It collects data and uses a Support Vector Machine with an RBF kernel to achieve a high accuracy of 99.1% for spammers and 99.9% for non-spammers.
* Paper [43] introduces SybilBelief, a semi-supervised learning framework for Sybil detection. SybilBelief outperforms existing methods, is resilient to label noise, and leverages Markov Random Fields and Loopy Belief Propagation for robust Sybil classification and ranking.

###### **Sangeeth Santhosh**

* In order to ensure that non-suspicious accounts are not mistakenly classified as suspicious, it is imperative to accurately identify the suspicious social media accounts for analysis. To classify something incorrectly would be to cast doubt on the honesty of a user. To achieve successful model evaluation, 5 research papers were reviewed.
* In [33], a system is proposed that uses sentiment analysis and machine learning techniques like Naive Bayes and Random Forest Classifier. In this research paper, the input chosen for classification only takes tweets that have been posted by accounts that were later suspended by Twitter. Thus, integrity of honest users is not compromised in this study.
* In [34], a novel method is proposed for identifying and evaluating phishing attacks on social networks. In this paper, the methods that are used are tested only on real data where the chances of false grouping is almost nil. This in turn means that only spam users are questioned and not honest users.
* In [30], examination and identification of opinion spam on social media is done with a special emphasis on Twitter. The methods that are proposed in this paper take into account only users who engage in opinion spam, meaning that integrity of legitimate users is not compromised.
* In [31], use of machine learning algorithms for spam filtering in bilingual tweets is touched upon using Roman Urdu tweets mainly. In this paper, each step of classification is done by domain experts thus considerably reducing chances of incorrect classification of legitimate users as spam.
* In [32], spam reporters and reportees are examined in the Twitter ecosystem using data mining techniques. In this approach, attributes of the reporter and reportee are focused upon which helps in the correct classification of spam users as spam and others as honest.

###### **Yeshwanth Reddy Chennur**

* Performed research using data mining and machine learning to detect suspicious activities on social media. Techniques like natural language processing and anomaly detection improve threat identification. These studies advance strategies for digital security.
* In the paper [25], The study proposes real-time detection of misinformation on Twitter. Advanced NLP and data mining will be used. Rapid detection is crucial to counter fast spread of false information.
* In the paper [23], The proposed framework integrates a mathematical model and decision-making to improve trust in social networks. It supplies a quantifiable trust metric to filter and rank content. This greatly advances a more reliable and safe online social setting.
* In the paper [35], It presents "Reveal," an online tool using machine learning to identify fake job ads in real-time. It protects job seekers from scams by distinguishing real and bogus listings. Reveal is beneficial for discerning authenticity of opportunities.
* The Paper [18], provides a hybrid deep learning-based technique for detecting suspicious flows in SDN, with better anomaly detection performance in social multimedia applications.
* The paper [24], provides a data-driven analysis of machine learning-based security dangers and remedies. It offers a comprehensive resource for professionals and academics working in this emerging field.

##### **DETAILED RESULTS OF ALL INDIVIDUAL GROUP MEMBERS**

###### **Krupaben Kothadia**

The reasons for focusing on data preprocessing and NLP as one of the tasks in our project is because they are crucial steps in developing a robust data mining and machine learning pipeline. Coming up with promising results rely on the good quality of the dataset. It’s very important to apply relevant NLP techniques to extract the crux of the data.

[An Integrated approach for Malicious Tweets detection using NLP](https://drive.google.com/file/d/1z9qYcbTsDz-IZOJNe5SgFdwLh9_qKNxi/view?usp=drive_link) [10]

* In order to properly protect user privacy, this study presents novel data preprocessing and NLP algorithms designed to analyze individual tweets rather than entire user accounts.
* The data was gathered from Twitter by the writers using the platform's APIs. The pre-classified tweets in this publicly accessible dataset were either labeled as spam or not, and the authors reevaluated the tweets that were incorrectly classified and connected to the same URLs. They used the text of tweets and their corresponding URLs to differentiate between these groups. The tweets that were gathered were arranged into a single file format.
* Additionally, the data was divided into tweets that contained URLs and those that did not. Data with URLs has been seen as spam, while data without them has been regarded as not spam.
* Next, the text and the URLs in the tweets were separated; these URLs served as one of the features. The author then processed the texts by eliminating stop words and using the separated words to determine the semantic meanings of the remaining words using sophisticated NLP techniques.
* This text has produced a number of features for the language model, one of which was utilized to determine the divergence ratio between the text's semantics and the page that could be accessed by clicking on the relevant URL.
* The study notes that this feature helps determine the divergence point (difference) between the posts, which ultimately determines whether or not something is different from the post itself. As a result, the authors have used this "divergence point" characteristic to create a new language model. The ability to categorize tweets as spam or not spam is a useful tool.

[Intelligent Analysis of Arabic Tweets for Detection of Suspicious Messages](https://drive.google.com/file/d/1AA1mL9x8Nbc39Pn1_JSlQLd4bKAZ32Vd/view?usp=drive_link) [12]

* Diacritics have a crucial role in language semantics in Arabic texts, helping to interpret word meanings. However, it is uncommon to use diacritics while writing Arabic language for social media, which makes Arabic tweet classification more difficult.
* The authors used Twitter's Streaming API, which gives developers access to tweet data. They used access tokens to gather tweets and Arabic keywords to extract a dataset of tweets. In order to accommodate Arabic characters, the acquired data was first converted to Microsoft Excel using UTF-8 encoding and stored in JSON format.
* Two crucial steps were engaged in the data preparation phase:
  + Data Filtration: Extraneous components were eliminated, including URLs, retweets, punctuation, and media material (pictures, videos), this procedure attempted to improve system performance and data quality while cutting down on training and testing time.
  + Data Tokenization: Because Arabic uses spaces to separate words and phrases, individual words in tweets can be tokenized. Spaces were used to separate each token, and tokens that were less than three letters long—usually conjunction words—were removed. Tokenization was essential to the stemming procedure.
* Then, two NLP approaches that reduce words to their base or root forms—stemming and lemmatization—were used. Because Arabic words have many different forms, lemmatization and stemming were essential for standardizing the lexicon, minimizing grammatical variances, and streamlining the indexing structure. As a result, categorization accuracy increased.
* The tweets were manually labeled into two categories: suspicious (label 1) and not suspicious (label 0). When these data sets are further feeded into models the performance is shown below Figure 4.1.1. [12 (Al-Ghamdi & Khan 2020)]

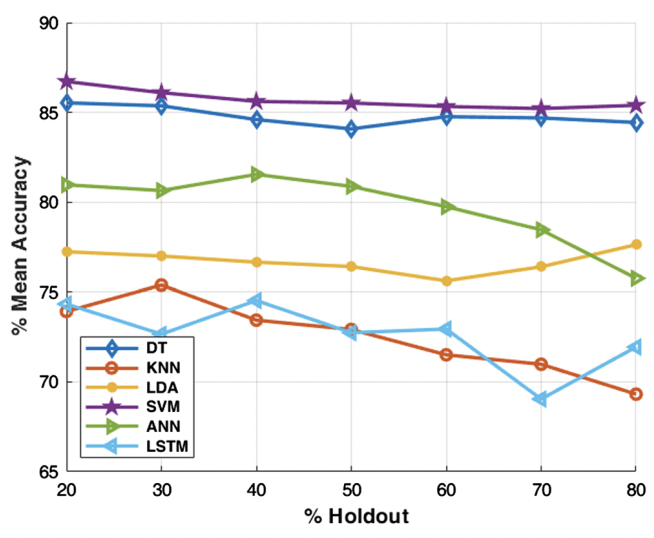


Figure 4.1.1 Mean Accuracy of Classifiers [12 (Al-Ghamdi & Khan 2020)]

* It is clear in Figure 4.1.1 [12 (Al-Ghamdi & Khan 2020)], that the SVM classifier has the best accuracy, followed by DT, while LSTM and KNN have inferior accuracy. When utilizing the bag-of-words model to categorize natural language, SVM is typically the first option. The high-dimensional textual data are mapped into high-dimensional feature spaces by SVM, which subsequently establishes the decision boundary.

[Mining user Message Pattern for Suspicious Behavior on Terrorism using NLP in Social Networks with Single Sign-On](https://drive.google.com/file/d/1STLuTC9mJx8LF8fihELth3KnG4n-0rLv/view?usp=drive_link) [13]

* The suggested method uses single sign-on authentication to monitor user behavior in real-time within the Gmail and Twitter databases. Twitter data contains timeline, sent, and received messages; Gmail data includes inbox and sent items. The IMAP protocol is used to retrieve Gmail messages.
* Suspicious user activity is extracted by the system framework. In order to collect user data, one must maintain data retrieval limits and log into Gmail and Twitter accounts using a single sign-on ID. Data retrieval is facilitated by the JavaMail API and Twitter4j API. Gmail accounts can be accessed over IMAP. Email images can be used for steganography to extract secret data using the LSB method.
* Next, the data is analyzed by NLP methods such as chunking, WordNet processing, spell checking, and POS tagging. WordNet processing finds semantic connections, chunking sentences structures the sentences, spell checking fixes misspelled words, and POS tagging identifies parts-of-speech in phrases. Figure 4.1.2 [13 (Selvan & Selvaraj 2017)] shows the tags that are used to create POS tags.

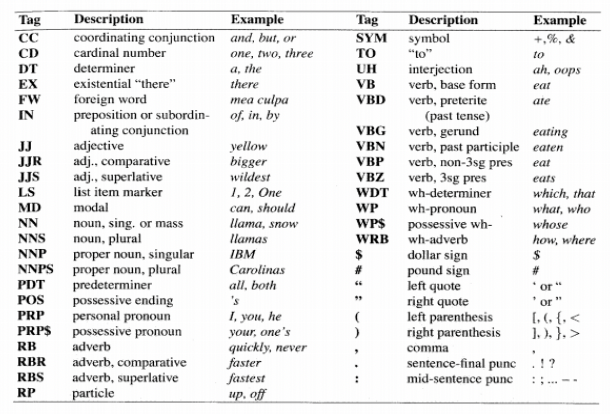


Figure 4.1.2 Part of Speech Tags [13 (Selvan & Selvaraj 2017)]

* Data extraction is guaranteed by the UDE (User Data Extraction) method, and the LSB algorithm retrieves data that has been hidden within images through steganography. The goal of NLP processing is to find information about terrorism in the content that has been retrieved.
* The NLP module processes data using Apache OpenNLP and employs methods to extract patterns from it. The modules that gave the dataset input receive these patterns back, allowing them to recognize suspicious user activity, particularly that connected to terrorism.

[Detection of User Cluster with Suspicious Activity in Online Social Networking Sites](https://drive.google.com/file/d/1-93xUJ5x4MJQEww1yoLOcA0i-wtr9g5z/view?usp=drive_link) [11]

* The goal of the authors' study is to provide a thorough method for examining communications on Social Networking Sites (SNS) in order to spot groups of suspicious individuals involved in suspicious actions. The authors set up a private experimental network called "Manipal Net." and it is based on gathering information from SNS through online surveillance, capturing crucial message information like sender, receiver, content, and timestamps.
* The authors use NLP and keyword-based systems to identify suspicious messages in this large dataset. Below is a summary of the various NLP systems that the authors have developed:
  + Sentiment Score Identification System: The words are rated according to their weight, in order to determine if a communication conveys a good or negative attitude.
  + Sentiment Count Identification System: This approach uses a list of 2230 positive and 3905 negative words to calculate the probability of positive and negative emotion word occurrence in a message.
  + Sentiment Identifying System Based on Training Sets: Based on the Sentiment Score Identification System results, this step determines which training set is best for classifying sentiment as positive or negative.
  + Dictionary-Based Topic Identifying System: This system makes use of NLP techniques to classify communications as normal or suspicious by using a topic dictionary constructed from subject-specific data to provide matching scores.
* In addition, the authors present Latent Semantic Analysis (LSA) as a tool for grouping people who use SNS to engage in deviant activity. Based on Singular Value Decomposition (SVD), LSA reveals hidden connections between words that don't seem to belong together, hence clusters of suspicious SNS users can be more easily identified.

[Suspicious Activity Detection of Twitter and Facebook using Sentimental Analysis](https://drive.google.com/file/d/1N9bz3ZE4Wi4g64_UIoMzsDgOplt54iCA/view?usp=drive_link) [14]

* In this study, 12,000 posts from 50 randomly chosen public Facebook profiles were included in one dataset, which contained 5,000 tweets from 50 randomly selected Twitter accounts.
* Data preprocessing requires a number of important processes when dealing with unstructured social media data. Before text normalization, which got rid of stop words, short words, and blanks, text cleaning removed extraneous text like HTML, emoticons, and hashtags. Using Part-Of-Speech (POS) tagging, text chunking divided text strings into related phrases and classified words as nouns, verbs, or adjectives. Text was categorized using Name Entity Recognition (NER) into predetermined classes.
* The cleansed data was saved in the.CSV format so that Valance Aware Dictionary Sentiment Reasoner (VADER), a lexicon-based technique, could be used for sentiment analysis. In order to categorize feelings as positive, negative, or neutral and to spot suspicious activity in Facebook and Twitter posts, VADER linked linguistic features to emotion intensities.
* The VADER scale had a range of -4.00 to +4.00, where 0 denoted neutrality. Hutto Normalization was used to normalize sentence-based characteristics, with an alpha of 15.

[Detecting malicious tweets in trending topics using a statistical analysis of language](https://drive.google.com/file/d/1pv0VpftrMVLTlXlObU3yEcF_XfkC8iGH/view?usp=drive_link) [43]

* It is possible to identify suspicious users on social networks by looking into their communication patterns. Links within messages, multimedia messages, and message themes are the three categories into which this study groups messages.
* Discussions about particular subjects are common, such as sports and TV shows and an unexpected shift in such topics or unrelated material could raise questions.
* Communications on some platforms can be marked with hashtags to help find anomalies. When mixed with other multimedia types, text, picture, audio, and video messages might reveal malicious intent. Content that is detrimental suggests a suspicious user, especially when used for blackmail. There is a chance that message links could be harmful.
* The paper introduces a brand-new approach built on NLP. NLP enables text analysis, with a focus on word tokenization and prediction in particular. N-gram models assess word probabilities inside phrases, whereas chain rule probabilities examine word sequences.
* Sequences can be compared to training datasets to prevent messages containing defined "bad words" from being sent. Maximum Likelihood Estimate (MLE) enhances the detection of malevolent users by computing word probabilities.

[Uncovering Cybercrimes in Social Media through Natural Language Processing](https://drive.google.com/file/d/1rAJ1djjueV3NpZ7rvedARvSnqdkh2rxB/view?usp=drive_link) [42]

* The article details a data preprocessing and NLP technique used to analyze tweets for a research study. To ensure the accuracy of the final model, the tweets were first cleaned to remove mentions, hashtags, and URLs. After that, letter casing was standardized using text normalization, and emojis were converted into their textual equivalents so they would not lose their meaning.
* Subsequently, the tweets were converted into the embedding language, typically English. Finally, each tweet was vectorized into a single vector that represented the average of all the word vectors in the tweet. This process was known as vectorization.
* With this approach, a matrix of tweets and their cosine distances was generated.
* A validation dataset was carefully chosen so that the performance of the similarity model in ranking related tweets could be assessed.
* Metrics like hits and discounted cumulative gain (DCG) were used to evaluate the model. DCG@K assessed the ranking's relevancy and order, while Hits@K ascertained whether a given tweet was among the top K most similar tweets.
* The study detailed the outcomes of applying this tactic to two protest situations in 2020: one in Colombia and one in the US. The TAGS application was utilized to gather the data, which came via Twitter. After Google News Embedding and word2vec were used for the embedding process, TinfoLeak was used to extract more data. Social network graphs were created using Gephi.

###### **Gautham Vijayaraj**

One of our important domains of research was secure data collection and privacy preservation. This is because finding the right balance between detecting suspicious activities in social media and respecting the user's privacy to maintain ethics, data confidentiality and integrity is crucial while dealing with real data. Thus, maintaining this balance is extremely important while searching for an appropriate data mining technique for our project.

[An Overview of Privacy Preserving Data Mining](https://drive.google.com/file/d/1XipNBU2_EKn2yklo7fJbQaWyXa1nXMy4/view?usp=drive_link) [5]

* This paper focuses on preserving privacy in data mining by exploring various methods to safeguard sensitive information. Data mining, while powerful in extracting substantial data, can also pose risks of identity theft and data insecurity.
* The primary objective is to identify algorithms that can ensure the confidentiality of private data during the mining process, necessitating the exclusion or modification of extracted data. The paper categorizes privacy preservation methods into data distribution, data distortion, data mining algorithms, data/rules hiding, and privacy protection.
* Data distribution methods are discussed in the context of centralized and distributed data, including vertical and horizontal data partitioning. Data distortion techniques like perturbation, blocking, aggregation, merging, swapping, and sampling are presented as means to modify data for privacy protection.
* Furthermore, the authors introduce evaluation criteria for privacy-preserving algorithms, including algorithm performance, data utility, privacy protection degree, and the complexity of data mining. Algorithm performance is assessed based on complexity and efficiency, with consideration for polynomial time versus exponential time algorithms.
* Data utility is a crucial concern to ensure that data modifications do not compromise its usability for analysis. Privacy protection degree measures the level of privacy assurance by accounting for the uncertainty introduced by hidden information. The paper acknowledges the challenges associated with different data mining techniques.

[Comprehensive Survey on Big Data Privacy Protection](https://drive.google.com/file/d/1lnxy_kihxdIVJHpah0283JycADq59Jwx/view?usp=drive_link) [6]

* The main focus area of this paper is to elaborate on the existing challenges and unresolved issues in privacy-preserving data mining (PPDM). A key issue of PPDM is how to manipulate data using a specific approach to enable the development of a good data mining model on modified data, thereby meeting a specified privacy need with minimum loss of information for the intended data analysis task.
* The current review study aims to utilize the tasks of data mining operations without risking the security of individuals’ sensitive information, particularly at the record level. The privacy preservation in the big data life cycle is considered at the data preprocessing and data mining task stages.
* Privacy Preservation In Data Providers:
  + Data providers cannot disclose any information because they regard their data as extremely sensitive and should never release private information because they are aware of the value of their data to data collectors. However, they may be willing to disclose some of their sensitive information for specific rewards, such as improved services or financial benefits.
  + If data providers cannot either block access to their sensitive information, then data collectors can misrepresent data collected by the data providers in such a manner that actual information cannot be easily revealed.
* Privacy Preservation In Data Collection:
  + The goal of preserving data privacy is to safeguard privacy during data collection and transmission to different data mining servers by finding the minimum portion of private information required to construct accurate data mining models
  + The direct disclosure of data to data miners will infringe on the privacy of data providers, particularly in cases where data miners execute mining algorithms using the data provided by data collectors and extract valuable information from data.
  + Thus, three types of approaches have been generally developed to conceal the raw data from their original value - data exchange, data cryptographic, and data modification.
* The other approaches to privacy preserving data mining techniques include data modification approaches like k-anonymity approach, l-diversity approach, and t-closeness approach. Table 4.2.1 [6 (Binjubeir et al. 2019)] shows the comparison of these approaches.

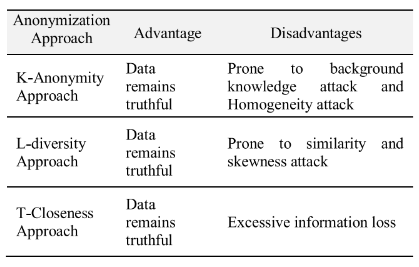
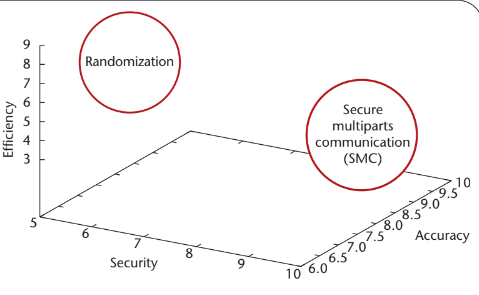


Table 4.2.1 A summary of data anonymization approaches [6 (Binjubeir et al. 2019)]

[Privacy-Preserving Data Mining: Why, How, and When](https://drive.google.com/file/d/1JF6NG6eQJ1_Y9zLhEyODuRKGJUtQuPMf/view?usp=drive_link) [7]

* The main focus area of this paper is to find how technology from the security community can change data mining for the better, providing all its benefits while still maintaining privacy. Data mining doesn’t inherently threaten privacy, and the paper focuses on two approaches that enable it without revealing private information: randomization and secure multiparty computation (SMC).
* The goal of data mining is to extract knowledge from data. Data mining is categorized into five tasks - Exploratory data analysis (EDA), Descriptive modeling, Predictive modeling: classification and regression, Discovering patterns and rules and Retrieval by content
* Most data mining applications operate under the assumption that all the data is available at a single central repository, called a data warehouse. This poses a huge privacy problem because violating only a single repository’s security exposes all the data. Whether a data warehouse is real or virtual is irrelevant: if the data mining algorithm can access the data, the opportunity exists for an attacker to get it, too.
* The following approaches to prevent disclosure of data from data mining:
  + **Data Perturbation** - Modifying the data so that it no longer represents real individuals. Uses Randomization to modify data extracted from datasets.
  + **Randomization** - Produces random samples from the set of data matrices satisfying the already discovered patterns or models.
  + **Secure multiparty computation** - Based on the idea that every piece of private information is validly known to one or more parties.
  + **Association rule mining** - Finds interesting associations and relationships among large sets of data items. Used in SCM approach to demonstrate protocol.
* Figure 4.2.1 [7 (Vaidya & Clifton 2004)] shows the efficiency, security and privacy scales of the approaches discussed.

##### Figure 4.2.1 Different approaches to privacy-preserving data mining and the trade-off between efficiency, security and accuracy [7 (Vaidya & Clifton 2004)]

[Comprehensive Survey on Big Data Privacy Protection](https://drive.google.com/file/d/1lnxy_kihxdIVJHpah0283JycADq59Jwx/view?usp=drive_link) [8]

* The main focus area of this paper is to discuss an approach for anomaly detection in Online Social Networks (OSNs) using data mining techniques. This analysis is carried out using Facebook dataset. It mentions profiling user social behavior by analyzing user updates, messages, clickstreams, photo uploads, posts, and comments.
* The discovery process consists of an iterative sequence of the following steps: Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation and Knowledge representation.
* DATA MINING APPROACHES TO ANOMALY DETECTION :
  + **Supervised Method** – treats the problem as a classification task with pre-labeled data, distinguishing between normal and anomalous observations.
  + **Semi-Supervised Method** – builds a model based on normal data to create a profile of normal activity. These methods work with both labeled and unlabeled data. They are useful when only a few instances of labeled normal data are available.
  + **Unsupervised Method** – trains an anomaly detection model using unlabeled data that includes both normal and abnormal instances. They are employed when labeled data with predefined labels like "anomalies" or "normal" are unavailable.

[Sentiments Analysis Of Twitter Data Using Data Mining](https://drive.google.com/file/d/1SGNHuQyCJ--drDbYgOgs12MNIApcoHFT/view?usp=drive_link) [9]

* The main focus area of this paper is to find an approach to analyze the sentiments of users using data mining classifiers. This paper presents a mechanism to predict the overall sentiments inclination of Indian people towards political situations and issues.
* After collecting raw tweets various preprocessing methods get applied to clean the data. The sentiment analysis is performed is using the following steps:
  + **Data Collection:** Training and testing tweets collected from twitter by using twitter searched API v 1.1 for various political leaders and parties in india.
  + **Preprocessing:** All userID,twitterId, userinfo from the tweets is removed. Duplicate tweets and retweets also get removed from the training dataset. After applying all cleaning methods text that is only with the tweeted text remains.
  + **Training DataSet and Testing Dataset:** For classifying a dataset, three classes, Positive, Negative, Neutral are used. To classify tweets into these categories, “SentiWordNet 3.0.0.” dictionary is used.
* Experimental results obtained demonstrate that k-nearest neighbor classifier gives very high predictive accuracy with an accuracy of 99.6456%

[Privacy Preservation of Social Network Users Against Attribute Inference Attacks via Malicious Data Mining](https://drive.google.com/file/d/1Gigpwllq6nNhbfu33l_-eNO7mxQF0CTZ/view?usp=drive_link) [40]

* The authors introduce a privacy-preserving technique called 3LP+ to address the growing concern of privacy in online social networks (OSNs). The key innovation in 3LP+ is its coordinated approach to preserving privacy across multiple attributes, avoiding conflicts that may arise when multiple runs of the algorithm are used independently.
* The 3LP+ technique is presented in three layers :
  + **Attribute Suppression (Layer 1):** In this layer, 3LP+ computes the sensitivity of each attribute and suggests to the user which attribute values to suppress. It generates sensitive rules and identifies the attributes that, when suppressed, would provide the most privacy protection. The decision to suppress an attribute ultimately rests with the user.
  + **Hiding Friendship Links (Layer 2):** If sensitive rules in Layer 1 still pose a threat (particularly those relying on link attributes), 3LP+ recommends hiding friendship links to disrupt the attacker's predictions. It selects links to hide strategically to minimize the impact on the user's social connections.
  + **Adding Friendship Links (Layer 3):** If sensitive rules involve link attributes that need to exceed a certain threshold, 3LP+ suggests adding new friends to the user's network. This helps make the rule inapplicable and further protects the user's privacy.
* The experimental results prove that the 3LP+ offers better privacy protection for users with multiple sensitive attributes compared to existing techniques.

[Intelligent Data Mining Technique of Social Media for Improving Health Care](https://drive.google.com/file/d/1-Sxd7wjjOK7qvZ-ep3tkkYA4QwF1nVhf/view?usp=drive_link) [44]

* The paper introduces an innovative approach to harness the power of social media for healthcare improvement, with a specific focus on cancer-related information. The primary objective of this research is to use intelligent data mining techniques to extract valuable insights from user-generated content on social media platforms, ultimately enhancing healthcare outcomes.
* Key components of this approach include: Data Collection, Text Processing, Pattern Taxonomy, Sentiment Analysis and Symptoms & Medication Identification.
* The results obtained from this approach go beyond traditional sentiment analysis by incorporating the identification of symptoms and their corresponding medications.
* The overall project requires data mining techniques which are efficient to detect suspicious activities. The approach discussed in this paper contributes to that goal as well and can hence be implemented.

###### **Avani Mundra**

One of the important aspects of social media data is its vulnerability to being unstructured. Several cutting-edge data mining and machine learning techniques have been adopted by researchers to handle such multimodal data and analyze suspicious activities in social media. This is crucial in ensuring information assurance and security in the vast social networks.

The following research papers have been reviewed to understand and study such techniques:

Cybercrime profiling: Text mining techniques to detect and predict criminal activities in microblog posts [16]

* The paper discusses text mining techniques to identify criminal activities in social media, specifically in microblog posts. It exploits the exclusivity of Twitter, which is the use of hashtags, to detect events and trending topics related to suspicious activities.
* The novelty lies in addressing two major issues which are data sparseness and the semantic gap when analyzing social media content for suspicious activities.
* The authors propose a data mining approach that involves three key steps: text corpus, corpus preprocessing, and classification process using similarity.
* The dataset finalized for this study consists of Twitter posts and includes 284 million following relationships, 3 million user profiles, and 50 million tweets, collected in May 2011. The first step involves cleaning this dataset by removing stop words and performing stemming. A special focus is given to the use of hashtags, which are defined as strings of characters preceded by the hash symbol (#). These are crucial as they are used to create communities around certain topics, thus helpful in analyzing suspicious activities.
* Once data preprocessing is done, The authors employ the Normalized Compression Distance (NCD) function to calculate a similarity distance between terms in the posts being analyzed and predefined suspicious terms that are already present in the database. If a post contains terms that exhibit similarity with those in their database, it is then classified as a suspicious post.
* To prove the efficiency of their method and the usefulness of the Normalised Compression Distance function, the authors illustrate with an example how a post titled a post containing the hashtag "#TerroristGroup" is declared suspicious since it contains threatening words such as ‘bomb’ and ‘explode’.
* The authors are thus able to achieve cybercrime profiling using this text mining technique and effectively address the issue of synonymy and polysemy. By introducing the disambiguation step. They further discuss their intention to enhance the system's execution time and precision in future research

An approach to detect abusive content incorporating Word2Vec and Multilayer Perceptron’ in depth [15]

* The paper discusses the detrimental effects of abusive content and proposes a novel model for detecting abusive text that employs the word2vec model and a compositional vector model to analyze text both semantically and syntactically.
* The research utilizes the English language dataset from Twitter and compares several machine learning and deep learning models to achieve best accuracy for suspicious content detection.
* The methodology consists of four main processes: preprocessing, word embedding, compositional vector model, and Multilayer Perceptron.
* The preprocessing step is employed to clean and transform the data. Following that, a word embedding layer is added using the word2vec model to represent tweets as word vectors. This step is followed by adding a compositional vector model to capture bi-gram information from tweets. A Multilayer Perceptron is added to these layers to classify the tweets containing hateful content.
* After comparing several models for embedding layer and classification layer as depicted in Fig 4.3.1 [16 (Alami & Elbeqqali (2015)] and Fig 4.3.2 [16 (Alami & Elbeqqali (2015)] given below, best results were achieved with word2vec model and MLP classifier, giving 86% accuracy.
* The future scope of this research entails incorporating emotion and sentiment features, domain-specific lexicons, and training the model to identify sarcasm.

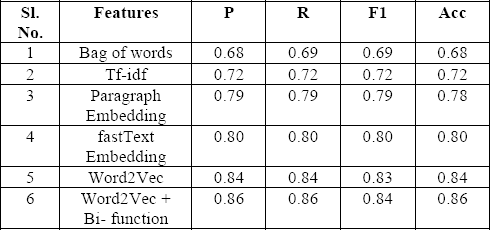
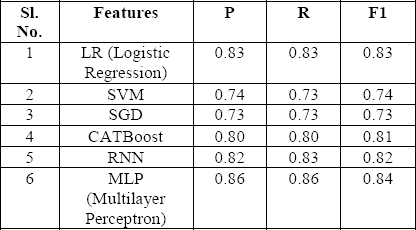
 

Table 4.3.1 Comparison of embedding layers Table 4.3.2 Comparison of classifiers

[16 (Alami & Elbeqqali (2015)] [16 (Alami & Elbeqqali (2015)]

A Framework to Detect and Prevent Cyberbullying from Social Media by Exploring Machine Learning Algorithms [17]

* The research paper addresses cyberbullying on social media, specifically Twitter, and proposes a prevention framework.
* An important aspect of ensuring security in social media is preventing cyberbullying. This research paper addresses the threats of cyberbullying and proposes a prevention framework using machine learning algorithms.
* The dataset involved contains English language tweets related to cyberbullying. This dataset has been preprocessed by removing usernames, emojis, and non-English words so that it is ready to be inserted into the model without creating noise or bias.
* Data annotation was carried out using the Valance Aware Dictionary and Sentiment Reasoner. The accuracy of the annotation was measured using kappa score, and achieved 0.83 on it.
* Once the data was preprocessed, two embeddinging layers, namely Term Frequency Inverse Document Frequency (TFIDF) and Bag of Words (BoW) were implemented and compared for accurate results.
* Post the embedding layer, the authors applied several machine learning algorithms including Logistic Regression, Multinomial Naive Bayes, Linear SVM, and Random Forest, which were employed for cyberbullying detection and classification.
* The best results were achieved with Random Forest with TF IDF embedding model, giving F1 score of 80.8% for cyberbullying identification and 58.4% for classification. This proved the superiority of the random forest model over other models.
* Upon further analysis, certain limitations of this research were also found, consisting of lack of large dataset and versatility of this model on other social media platforms. The authors intend to take this up as future scope of this project, along with exploring other advanced NLP techniques.

Social Big Data Mining Framework for Extremist Content Detection in Social Networks [38]

* This study analyzes the importance of social networks like Facebook as rich sources of user generated text data for expressing opinions, hence a source for detecting extremist and suspicious content that are a threat to society.
* The framework proposed consists of four main steps: data extraction, sentiment analysis, extremist content detection, and data storage in HBase for further analysis.
* The streaming data is extracted from Facebook public pages. This is carried out by sending streaming requests to Facebook API and collecting the data in JSON format, which is then parsed to extract filtered and raw data. Authentication is performed through OAuth for access to the Facebook Social Graph.
* The next step is sentiment analysis. This is broken down into six main steps which are: Entity extraction and classification, aspect extraction and categorization, opinion holder extraction and categorization, time extraction and standardization, aspect sentiment classification, opinion quintuple generation.
* For the extremist content detection, a second layer of processing os added that contains three behavioral linguistic identifiers: Leakage - which indicates an intent to harm and alerts on research, planning, and execution of violent acts, Fixation, where the person spends a lot of time studying the believed enemy and Identification, in which the subject identifies with previous attackers or radicals and may exhibit narcissism and fantasies.
* The analyzed data is then stored in HBase and communication is established through Representational State Transfer (REST) calls.
* In order to provide a user interface for the detection framework , a front-end in JavaScript is created, connected to a Java framework, and linked to HBase for data retrieval and analysis. This is efficient in providing a graphical representation of the data. The entire pipeline is depicted in Fig 4.3.1 [38 (Mouhssine & Khalid (2018)].
* The outcome of this semi automated framework is to ensure that extremist content and terrorist activities are detected before they occur ensuring safety of the citizens worldwide.

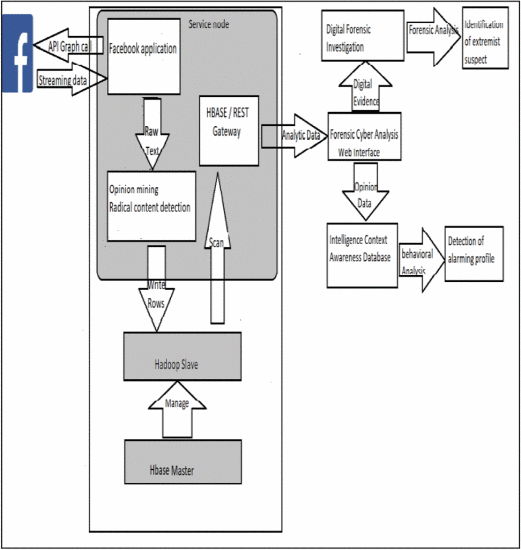
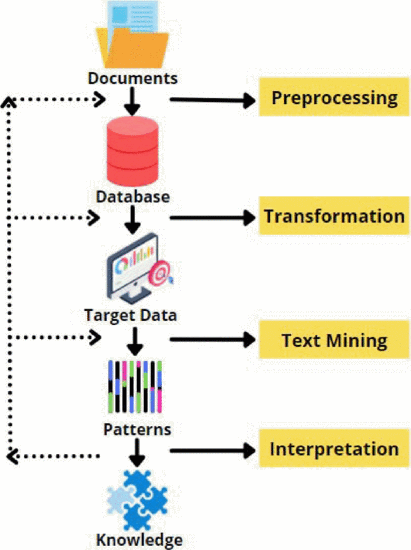


Fig 4.3.1 Extremist Content Detection Pipeline [38 (Mouhssine & Khalid (2018)]

Predicting Abnormal User Behaviour Patterns in Social Media Platforms based on Process Mining[39]

* The research study highlights the aggravating issue of cyberbullying in a socially connected world and proposes to tackle the same using process mining, which combines data science with management systems for evaluating organizational functions using log data.
* The major aim of the proposed solution is to analyze Twitter tweets live time detecting cyberbullying and further provide the user time-based metrics.
* Text mining has been performed on the dataset containing 3 attributes (Serial Number, Tweets, Text Label) has1064 records. This is used to extract meaningful textual data from tweets and Schizophrenic Discourse Framework has been employed to achieve text mining.
* The data is then preprocessed by performing data cleaning, data transformation, feature extraction, tokenization, stop words removal, data compression, data discretization and data normalization. This process is shown in Figure 4.3.2 [39 (S. G. et al. (2023)].
* A Recurrent Neural Network model has been built on the dataset and various classifiers such as Naïve Bayes, Decision Tree, Maxent, and Support Vector Machine, are evaluated for their performance in cyberbullying detection.
* To evaluate the model, several techniques such as precision, recall and F1 score have been used. It is evident from results that Naive Bayes classifier gives best results in cyberbullying content detection with accuracy of 0.725 with a loss 0.544.
* The proposed study thus successfully captures abusive content in social media by adopting the process mining technique and is hopeful of continuing this study by incorporating visual media in future research.



4.3.2 Text Mining process [39 (S. G. et al. (2023)]

###### **Justin Young**

Machine learning can be a powerful tool when it comes to ensuring security within malicious activity models by detecting and classifying attacks in various different environments such as mobile devices, web-based systems and the cloud.

Behavior-Based Malware Detection System Approach For Mobile Security Using Machine Learning [28]

* This paper covers the expanding mobile ecosystem, and the security concerns that come with that. This study proposes a process to design a behavior-based malware detection system for mobile devices.

****

Figure. 4.4.1 System Design Flow [28 (Vanjire & Lakshmi 2021)]

* This process consists of 4 steps: begin monitoring the system on application startup, show application usage, designate ignored applications and hide applications Figure. 4.4.1 [28 (Vanjire & Lakshmi 2021)].
* For this experiment, the K-Nearest Neighbor, Naive Bayes, and decision tree machine learning algorithms are compared in a performance analysis on the “Derbin” dataset in Py-charm.
  + An Android phone is used for testing, as it is the most popular operating system in the mobile phone industry which allows for greater security risks.

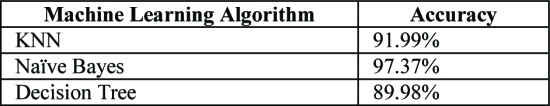


Table. 4.4.1 Result Accuracy [28 (Vanjire & Lakshmi 2021)]

* The analysis concludes that the Naive Bayes algorithm was most accurate in detecting and classifying malicious behavior in the defined mobile environment Table. 4.4.1 [28 (Vanjire & Lakshmi 2021)].
* Other research outcomes from this paper include the preference of dynamic malware detection and feature matching as a weakness for machine learning-based models against more sophisticated attacks

Comparative Analysis of Machine Learning Models in Computer Network Intrusion Detection [26]

* This paper aims to provide a general comparative of machine learning algorithms for intrusion detection on the CICIDS 2017 dataset.
* A supervised machine learning approach is proposed for a binary classification of network traffic as either malicious or benign

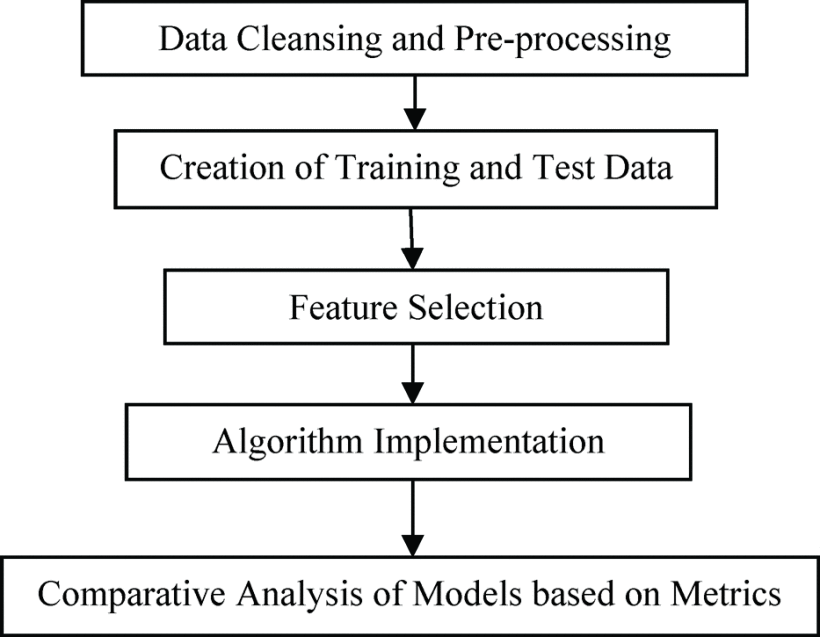


Figure. 4.4.2 Methodology [26 (Osa & Oghenevbaire 2021)]

* The procedure involves reducing the size of the dataset by removing null or incorrect values, splitting the dataset and implementing a Random Forest Regressor technique to narrow down features.
* The algorithms XG-Boost, Decision Tree, Random Forest, K-Nearest Neighbor (KNN), Multi-Layer Perceptron (MLP) and Quadratic Discriminant Analysis (QDA) are chosen for comparison.

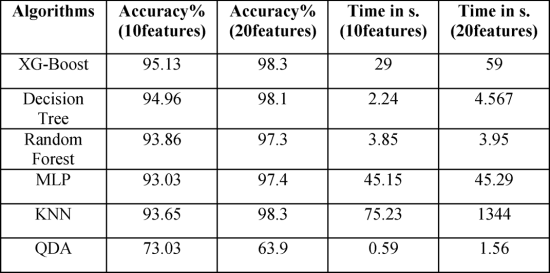


Table. 4.4.2 Result Summary [26 (Osa & Oghenevbaire 2021)]

* The study concludes that XG-Boost was most accurate in both experiments, while Random Forest had a good combination of high accuracy with low training time. Additionally, QDA had very low training time yet had the lowest accuracy scores among the group.
* This research provides valuable information regarding the performance of machine learning algorithms that may be considered when implementing a social media security model.

Cooperative Machine Learning Techniques for Cloud Intrusion Detection [27]

* In this paper, a new network intrusion detection model Secure Packet Classifier (SPC) is proposed for cloud intrusion detection.

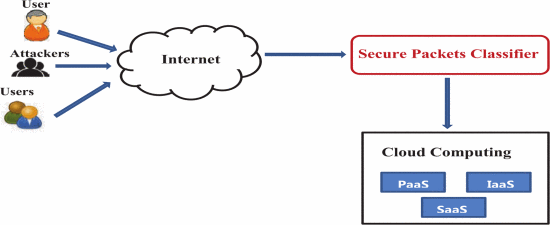


Figure. 4.4.4 The proposed model. [27 (Chkirbene et al. 2021)]

* This model aims to detect and classify cloud anomalies using collaborative filtering techniques between two machine learning algorithms.
* The framework for this model consists of three steps:
  + 1. Off-line data processing (partition the dataset with optimized running times and performance metrics)
  + 2. Train the model (ML model creation, confusion cube construction, SPC model creation Figure. 4.4.4 [27 (Chkirbene et al. 2021)])
  + 3. Test the model (verify SPC performance with respect to other machine learning algorithms)



Figure. 4.4.5 Model creation using UNSW dataset. [27 (Chkirbene et al. 2021)]

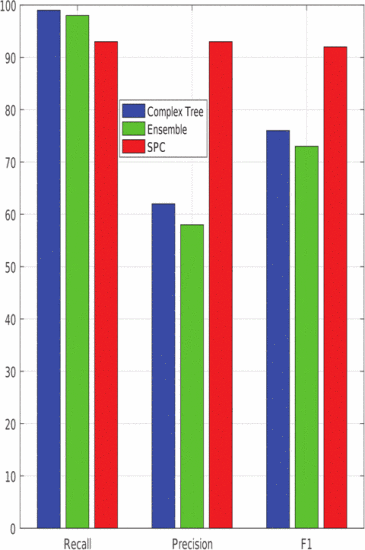
****

Fig. 4.4.6 Recall, precision and F1 of SPC model compared to complex decision and ensemble boosted tree. [27 (Chkirbene et al. 2021)]

* This study concludes that the proposed SPC model can detect and classify cloud anomalies with an accuracy of 81%, which is an improvement of 20% over traditional approaches. Figure. 4.4.6 [27 (Chkirbene et al. 2021)].

Ensuring Anomaly-Aware Security Model for Dynamic Cloud Environment using Transfer Learning [37]

* This paper focuses on introducing a new concept “transfer learning” to detect malicious activity on the cloud.
* This process involves improving learner performance by training on different attack types to detect unseen variants of new attacks.
  + Two different transfer schemes allow the model to support multiple types of attack classification and detection.
  + Signatures and features with samples in the target domain are matched, focusing on the unknown samples.
* This model was proven to be highly effective in classifying known attack types and detecting unknown attacks in the cloud when paired with a deep learning model.
* This study provides a powerful technique that can be used in cloud based social media systems.

Machine Learning-based web security intrusion detection system [29]

* This paper identifies SQL-based attacks and XSS cross site scripting as principle vulnerabilities in web security intrusion detection systems, and aims to compare the performance of machine learning models against these attacks.
* The models chosen for analysis are Adaboost, AVM, Decision Tree, Random Forest, Logistic Stiff Regression, KNN and Bayesian.

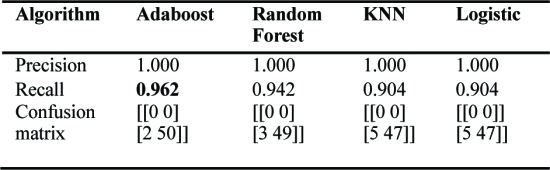


Table 4.4.3 Performance comparison of seven models-1 [29 (Chen et al. 2021)]

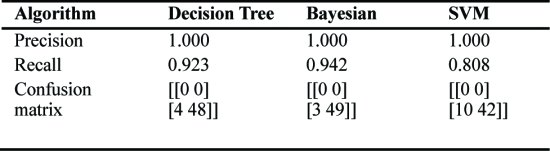
****

Table 4.4.4 Performance comparison of seven models-2 [29 (Chen et al. 2021)]

* The results of the first experiment on SQL-based attacks indicate that while all seven algorithms had a perfect rate, the Adaboost algorithm had the highest accuracy rate with a stronger recall score of 0.962 Table. 4.4.3 [29 (Chen et al. 2021)]
* The second experiment is a text disambiguation process on samples using a defined text classification on XSS data.
  + Aspects such as word vector dimensionality, words per entry, neural network batch size, neurons per model are reduced in capacity for model optimization.
* The results of this experiment indicate that the SVM model had the highest detection and accuracy rate of the models tested, however had a high number of false positives and lacked sufficient learning capability with average data containing large numbers of nun-UNK words.

###### **Anuranjan Dubey**

Machine learning plays a critical role in security and information assurance by using data-driven algorithms to identify and mitigate threats, detect anomalies, and protect sensitive information, making it an indispensable tool in the ever-evolving landscape of cybersecurity.

[Cyberthreat Detection from Twitter using Deep Neural Networks](https://drive.google.com/file/d/1NusmoTBnGaOwuXmbHVApEJaraEF95o0H/view?usp=drive_link) [22]

* The study focuses on the utilization of machine learning algorithms to enhance the accuracy and efficiency of tools designed for classifying and categorizing security-related content on the Twitter platform.
* The proposed method employs deep neural networks, specifically Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) networks, which have demonstrated superior performance compared to other established approaches. The evaluation of this tool indicates an impressive average true positive rate of 94% for classification and an average F1-score of 92% for named entity recognition.

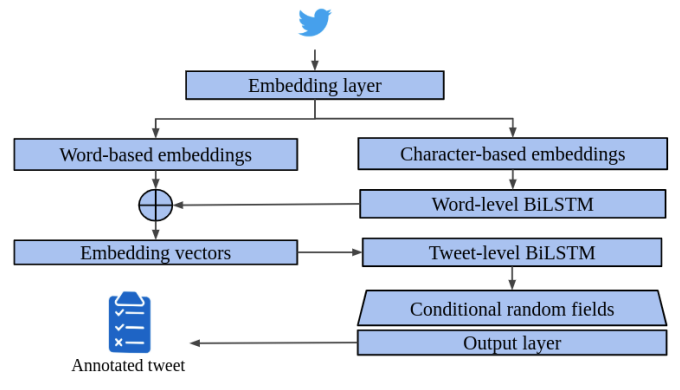


Figure. 4.5.1: Bidirectional LSTM architecture for named entity recognition. [22 (Dionísio et al. 2019)]

* This method leverages deep neural networks to streamline the process of gathering threat intelligence for security professionals, eliminating the need for manually sifting through a large volume of irrelevant data.
* The tool consists of a well-defined processing pipeline that systematically analyzes cybersecurity information obtained from Twitter. It incorporates a binary classifier based on CNN to distinguish tweets containing pertinent security-related data, thus separating valuable information from irrelevant content. Figure. 4.5.1 [22 (Dionísio et al. 2019)]
* Additionally, it utilizes a Named Entity Recognition (NER) model implemented as a Bidirectional Long Short-Term Memory (BiLSTM) neural network, extracting named entities such as indicators of compromise from relevant tweets. This extraction is facilitated through the embedding layer of the network, encompassing both word-level and tweet-level BiLSTM.
* The automation of the cybersecurity information analysis and summarization process eliminates the tedious task of manual data curation, providing security analysts with synthesized knowledge.
* The deep neural networks embedded within this tool play a pivotal role in identifying relevant information and extracting valuable entities. This synthesized knowledge becomes a valuable resource for Security Operations Centers (SOCs) aiming to effectively monitor and secure IT infrastructures.
* In summary, this proposed tool significantly enhances the efficiency and effectiveness of obtaining threat intelligence by alleviating the need for manual data sifting.
* The effectiveness of this method is validated through simulation scenarios, where it excels in classifying relevant tweets and extracting named entities.
* Real-world implementation with empirical data consistently demonstrates high accuracy in tweet classification and named entity recognition, underscoring its practical value for security analysts and professionals.

[CyberTwitter: Using Twitter to generate alerts for Cybersecurity Threats and Vulnerabilities](https://drive.google.com/file/d/14LY8rxklXibD4hygSqNHTrHp7BaC_91m/view?usp=drive_link) [19]

* The tool, known as SVCE (Security Vulnerability Concept Extractor), is employed to analyze collected data, extracting conceptual entities related to cybersecurity vulnerabilities from Twitter. This process aids in filtering out irrelevant information, focusing on critical security-related content.
* Analysts create system profiles, incorporating software and hardware information. These profiles are generated automatically using a tool called Facter, a cross-platform Ruby library designed for retrieving system data.
* SWRL (Semantic Web Rule Language) is a language utilized to express rules and logic within the context of the Semantic Web. It enhances unstructured text by adding semantic meaning, converting it into structured RDF data for knowledge graph construction.
* RDF Triples (Resource Description Framework) serve as a data model for structuring web-based information using triples. These triples extract structured information, particularly named entities, and link them to URIs. This preprocessing step facilitates text data integration into knowledge graphs and linked data. [19 (Mittal et al. 2016)]
* Named Entity Recognition (NER) is a natural language processing technique for identifying and categorizing named entities (e.g., people, organizations) in text. It plays a crucial role in annotating unstructured text with structured data, enabling knowledge graph construction and other semantic web applications.
* CyberTwitter generates personalized alerts based on a real-time knowledge base that updates according to a user's system profile. Intelligence extracted from Twitter is represented as RDF triples in a semantic knowledge base, and the tool employs SWRL rules and an intelligence ontology to reason with system profiles and knowledge base data, generating pertinent threat alerts for human review [19 (Mittal et al. 2016)].

| Figure. 4.5.2: Graphical representation of RDF for tweets by SVCE. [19(Mittal et al. 2016)] | Figure 4.5.3: A framework for monitoring and analyzing tweets related to cyber attacks. [19 (Mittal et al. 2016)] |
| --- | --- |

* The paper concludes by mentioning the use of semantic textual similarity systems to enhance tweet selection, recognizing relevant tweets based on their content, and expanding the word embedding model, particularly in the context of cybersecurity text.
* The performance of the CyberTwitter system was assessed through a ten-day experiment, and the majority of alerts proved valuable. This experiment demonstrates the tool's effectiveness in extracting real-time cybersecurity insights from Twitter data and its ability to issue timely threat alerts.

[Weakly Supervised Extraction of Computer Security Events from Twitter](https://drive.google.com/file/d/1FfclqLsECMhNk0dr1bzHV5BQhCBrWNvJ/view?usp=drive_link) [20]

* The paper compared baseline algorithms, such as One-Class SVMs and Expectation-Maximization (EM), with the latter exhibiting superior precision and recall in the context of security event detection on Twitter.
* Beyond the baseline methods, the study also employed specific algorithms, particularly in the areas of Natural Language Processing (NLP) and entity extraction. The proposed approach, Expectation Regularization, outperformed the baseline methods.
* The process of collecting candidate events involves gathering a collection of tweets related to specific seed events. Seed instances, represented as historical examples (Entity, Date), are used to train a system for the real-time detection of similar events on Twitter.
* Extracting candidate events comprises steps like providing historical seed examples, collecting tweets mentioning seed events, monitoring relevant keywords (e.g., "hacked" for account hijacking), extracting named entities and tweet dates, and filtering candidate events based on seed examples.
* The paper's results indicated that the weakly supervised approach it proposed surpassed previous methods in terms of precision and recall for security-related events, although specific accuracy percentages were not mentioned.
* Label regularization and other techniques were employed to address challenges related to learning from limited positive seeds and unlabeled events.
* The paper concluded by highlighting Twitter's value as a resource for security event information and emphasizing the effectiveness of the weakly supervised seed-based approach.
* The proposed weakly supervised approach demonstrated its superiority over previous methods, including heuristic negatives, semi-supervised EM, and one-class SVMs. Significant enhancements in precision and recall were observed, particularly for security-related events like DoS attacks, data breaches, and account hijacking.
* The weakly supervised seed-based approach allows for the rapid training of event extractors with minimal supervision, offering promise in the realm of security content classification on Twitter.

[An effective security alert mechanism for real-time phishing tweet detection on Twitter](https://drive.google.com/file/d/1b6pghiZ9c4T-rlq8Gq_CX7Q2HIabxIC9/view?usp=drive_link) [21]

* Feature extraction played a pivotal role in the success of the security alert mechanism, utilizing a comprehensive set of 22 features spanning various categories, complemented by 7 additional features identified by the WEKA tool. This robust feature set significantly improved the mechanism's ability to distinguish between phishing and legitimate tweets.
* The WEKA machine learning tool was instrumental in both feature selection and evaluation. By incorporating the 7 additional features recommended by WEKA, the mechanism enhanced its capacity to differentiate between phishing and legitimate tweets.
* During machine learning training, a dataset containing 2973 training data with URLs from Twitter, labeled as phishing or safe, was employed.
* The mechanism's real-world implementation demonstrated its efficiency in live scenarios, providing timely alerts to Twitter users regarding potential phishing URLs, a critical capability given Twitter's rapid information dissemination.
* The choice of the Random Forest machine learning technique contributed to the mechanism's success, owing to RF's robustness and its ability to handle complex datasets, making it well-suited for phishing detection on a dynamic platform like Twitter.
* While the security alert mechanism primarily detects phishing URLs, it also considers other features related to tweet content, user information, and network properties. Features like URL length, SSL connection, Hexadecimal, Alexa rank, Age of domain - Year, Equal, Digit in host, Host length, Path length, Registrar, and Number of dots in host were used in the final evaluation.
* The security alert mechanism based on Random Forest demonstrated impressive performance in simulating phishing tweet detection, achieving a precision of 94.64% and a recall of 95.49%.
* When implemented and evaluated using real Twitter and PhishTank datasets, the mechanism exhibited outstanding accuracy, with a rate of 97.50% in swiftly alerting users to potential phishing URLs.

[A feature-based approach to detect fake profiles in Twitter](https://drive.google.com/file/d/1TWzY3eHDCt7UNTkeZWDrVlzAeQJCDo-G/view?usp=drive_link) [36]

* A feature-based approach for fake Twitter account detection used 24 distinct features.
* Three machine learning algorithms, Logistic Regression, Support Vector Machines (SVM), and Random Forest, were evaluated, with Random Forest demonstrating superior performance.
* Mentioned data mining tools included DeepScan, SybilBlind, Botometer, and COLOR+, each serving various purposes.
* Methodology involved data preprocessing, supervised machine learning, and an 80:20 training-to-testing data split Figure. 4.5.4 [36 (Kaubiyal et al. 2019)]..

| Figure 4.5.4: Architecture of the proposed model. [36 (Kaubiyal et al. 2019)]. | Table 4.5.1: Results for supervised machine learning algorithms. [36 (Kaubiyal et al. 2019)]. |
| --- | --- |

* The primary goal was combating fake accounts and automated bots on social media platforms.
* The 24 features covered categories like account-based, tweet-based, ownership-detail-based, and URL-based characteristics.
* The study didn't specify the generalizability of the approach to other social media platforms.
* Evaluation showed Random Forest outperformed other algorithms with 97.9% accuracy Table 4.5.1 [36 (Kaubiyal et al. 2019)].
* Future work includes studying human-operated fake accounts, enhancing bot detection, incorporating sentiment analysis, and developing real-time bot identification tools.

###### **Rahul Nayak**

###### 

###### Before we use machine learning to detect and address the many suspicious activities on social media, we must understand the different types of suspicious activities that can happen there. To protect the digital world, we must first identify and understand the different types of suspicious activities that can happen on social media.

Suspicious Behavior Detection: Current Trends and Future Directions [4]

* In Paper [4], a range of suspicious activities are explored, including social spam, social Sybils, and link farming. These activities can have detrimental effects, such as misinformation and trust erosion. Detecting and addressing them is essential for the integrity and security of online social environments.

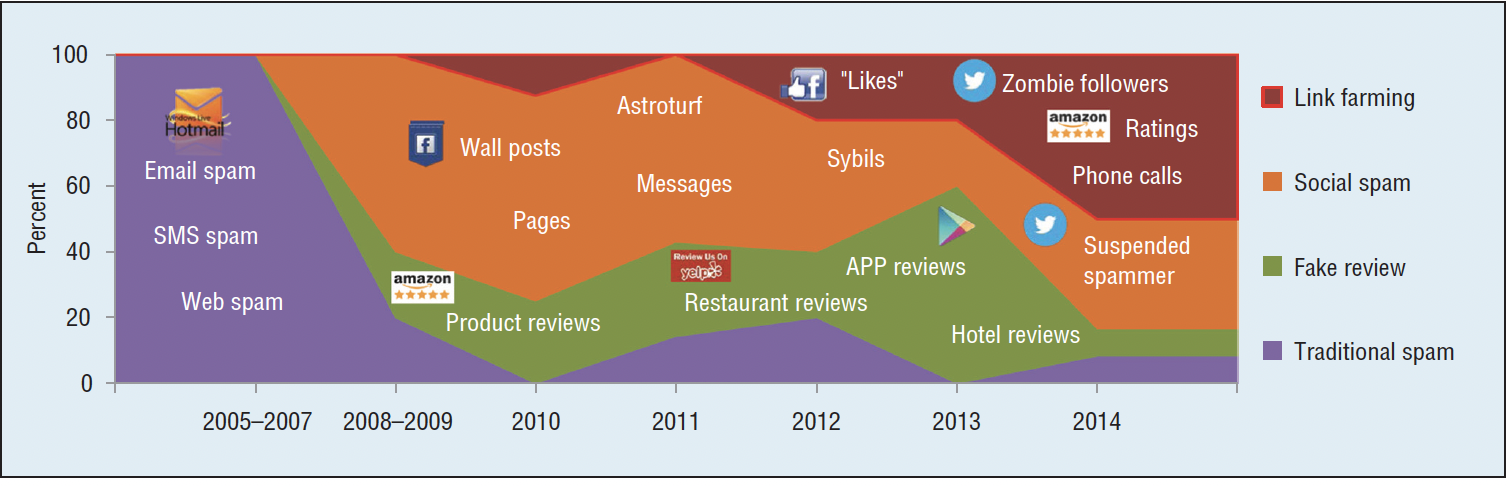


Figure 4.6.1: Types of Suspicious activities on social media [4 (Jiang and Peng Cui 2016)]

In this paper various types of suspicious activities are discussed, they are as follows

* Social spam is the unwanted distribution of irrelevant, deceptive, or malicious content on social media platforms, messaging apps, and online communities. It can take many forms, including comment spam, message spam, fake accounts, and content spam. There are various types of spam one can get to see on social media
  + Comment spam: comments in discussion threads and comment sections that advertise goods, services, or connections to dubious websites are known as comment spam.
  + Message spam: unsolicited communications in group and private conversations that lead to malicious websites or involve scams or phishing attempts are referred to as message spam.
  + Fake accounts: By seeming to be real people or organizations, fake or fraudulent profiles spread spammy information and make it harder for users to tell the difference between spam and authentic accounts
  + Content spam is defined as bulk publication of irrelevant, repetitive, or low-quality content. Examples of this type of content include clickbait articles, adverts, and connections to other websites.
* Social Sybils are a type of fake account on social media that is designed to deceive users into thinking they are interacting with a real person. These accounts are often created and used by malicious actors to spread propaganda, manipulate public opinion
  + Social Sybils can be very effective in their deception because they are often very well-crafted. They may use real names, profile pictures, and other personal information to create the illusion of authenticity. However, there are a few key signs that can help you identify a Social Sybil account:
    - They are often very new to the platform. Social Sybils are often created quickly and in large numbers, so they may have very few followers or friends.
    - They post a lot of spammy content. Social Sybils may post large volumes of repetitive or irrelevant content, such as ads, clickbait articles, or links to malicious websites.
    - They engage in unusual behavior. Social Sybils may follow or unfollow users in large numbers, or they may send unsolicited messages.
    - They impersonate real people or organizations. Social Sybils may use the names, profile pictures, or other personal information of real people or organizations to create the illusion of authenticity.
* Link farming is a deceptive technique to manipulate search engine rankings, creating an interconnected web of websites to boost the visibility of specific pages or websites.Some of the types of Link farming as surveyed according to paper[4] are:
  + Systematic interconnected networks: Here,Automated tools or manual methods are used to construct a web of interconnected sites.
  + Search engine manipulation: In this type of Link farming it aims to deceive search engines into thinking that the target web pages are authoritative and relevant.
  + Propagation of false or low-quality content:Here, Link farming often spreads misleading, false, or low-quality content. These websites may lack genuine informational value and only exist to improve the search rankings of associated pages.
  + Undermining content reliability: Link farming undermines the reliability of online content by prioritizing pages based on artificial link networks rather than the credibility or authenticity of the information presented.

Machine Learning Based Twitter Spam Account Detection: A Review [2]

* Paper [2] delves into spam detection on Twitter. It highlights the negative impact of spam on the platform and presents a framework for spam detection using machine learning, with a training phase and a testing phase. Various classifiers are employed, and model performance is evaluated using metrics such as accuracy, detection rate, and precision.
  + Twitter, a popular microblogging platform with a large user base, is particularly vulnerable to spam attacks. Spammers use social engineering techniques to send spam tweets, spam links, and other malicious content. This has led to an increase in the number of fake spam accounts on Twitter, which harm regular users by compromising their security and privacy.
  + The authors Shivangi and Rakesh [2] discuss how machine learning (ML) can be used to detect spam accounts on Twitter. ML allows us to create models that can learn and make predictions. In this context, spam detection is a binary classification problem, where the goal is to classify Twitter accounts as either spam or non-spam. A spam detection model typically has two phases: training and testing. In the training phase, the model is taught to identify spam accounts using labeled samples. In the testing phase, the model is used to classify unlabeled samples as spam or non-spam.
  + The authors divide the features used for spam account detection into three categories: content-based, user-based, and graph-based features. Content-based features include the number of URLs in a tweet, the number of characters in a tweet, and whether or not the account has a profile picture. User-based features include account characteristics such as the number of followers and followings. Graph-based features use the social network structure to identify spam accounts. The authors emphasize that choosing the right features is essential for effective detection.

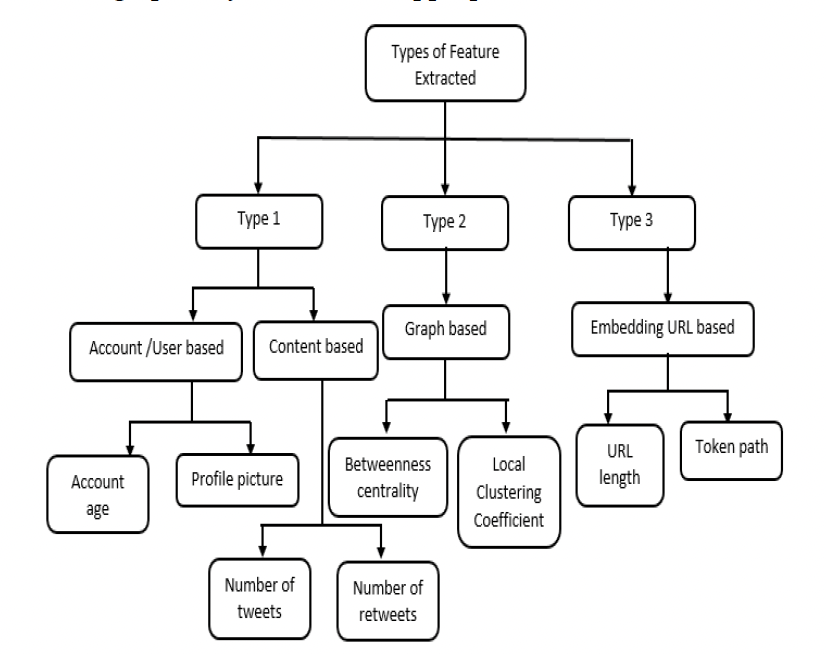


Figure 4.6.2: Flowchart depicting types of features extracted from twitter[2 Shivangi and Rakesh (2018)]

* + The authors identify several challenges in spam detection, including spammers fabricating features, the imbalance between spam and non-spam accounts, and spam drift, where spammers change tactics to avoid detection. Despite advances in machine learning (ML)-based spam detection, achieving 100% accuracy remains a challenge due to spammers' evolving evasion tactics.

Malicious Profile Detection on Social Media: A Survey Paper [3]:

* Paper [3] by Shruti Shinde and Dr. Sunil B. Mane addresses the critical issue of identifying and mitigating malicious profiles on social media platforms. Malicious profiles, such as fake accounts, spam accounts, bots, and more, can have detrimental consequences for individuals and the online community as a whole.These are some of the variety of malicious profiles that are discussed in the paper.
  + Fake profiles: These accounts are created with the intent to deceive others by impersonating real users or distributing misleading information. For example, a fake profile might be used to spread misinformation about a political candidate or to scam people into sending money.
  + Spam accounts
  + Bot accounts: These are automated accounts that are designed to artificially amplify certain content, manipulate metrics, or engage in other activities that compromise the integrity of the platform. For example, a bot account might be used to like and retweet a post on social media in order to make it appear more popular than it really is.
  + Clone accounts: These accounts duplicate another user's information, often for fraudulent purposes. For example, a clone account might be used to impersonate a celebrity in order to scam people into sending money or revealing personal information.
  + For most of the research conducted in this domain data is collected from social media platforms like Twitter, Facebook, and LinkedIn. They gather user account information and posts.
  + Data collection often relies on platform APIs, but privacy policies and restrictions can limit the availability of certain data.
  + Different features or attributes are selected from the dataset to create the feature set for machine learning models.
  + These features include information about user profiles (e.g., name, followers, posts), content posted by users, and other relevant attributes.
  + Researchers use various supervised and unsupervised machine learning algorithms to classify profiles as malicious or genuine.
  + In the paper [3] ,classification algorithms that are used include Support Vector Machines (SVM), Naive Bayes, Random Forest, Decision Trees, Neural Networks, and K-nearest neighbors.
  + The results that the authors get vary but often fall within the range of 80-95%. For instance, a study on LinkedIn reported an accuracy rate of 84%, while a Twitter study achieved 90.40% accuracy in identifying fake accounts.
  + The results are obtained by comparing predicted labels (fake/genuine) to actual labels, typically using a confusion matrix.

Detecting spammers on social networks [1]

* Paper [1] focuses on detecting spam accounts on Sina Weibo, a Chinese social network. Data collection methods and user-based feature analysis are detailed. A Support Vector Machine with an RBF kernel is used for detection, with high accuracy achieved.
  + The authors of this paper collected a dataset from Sina Weibo, including over 30,000 users and more than 16 million messages.
  + The authors began by selecting 100 regular users and 50 spammers as their data sources. They chose these users from a variety of categories, such as celebrities, businesses, and government accounts, to ensure the dataset's diversity.
  + They created two types of data crawlers, one for regular users and one for spammers. The regular user crawler extracted the list of regular users' followers, as regular users are less likely to follow spammers. The spammer crawler extracted the list of spammers who engage in specific malicious behavior.
  + Using the data crawlers, the authors collected information from 30,116 Weibo users, allowing them to extract a significant amount of data for their analysis.
  + For each user in the dataset, the authors crawled information within the 500 most recent messages. This information included the number of followers, followers, days since account creation, and various message attributes such as the number of reposts, comments, and likes.
  + The paper analyzes the collected data in-depth, focusing on content-based and user-based features to distinguish between spammers and non-spammers.
    - Content-based features:The authors randomly selected 300 spam and 300 non-spam messages, each assigned a unique integer identity value. They examined three key aspects of spam messages: the distribution of repost numbers, the number of comments, and the number of likes. They observed that most spam messages have low repost, comment, and like counts, indicating that they generally receive less engagement from users.
    - The authors also analyzed the presence of mentions in messages, the inclusion of URLs, and the use of hashtags. They found that most spam messages do not contain mentions and tend to include URLs for advertising purposes.
    - User-based features:The authors also analyzed user-based features to differentiate spammers from non-spammers. They used cumulative distribution functions (CDFs) to examine various attributes, such as the number of followers, the number of followers, the fraction of followers per follower, and the number of created days. They also considered the number of messages posted per day and the average number of URLs in messages as key user-based features.
  + The core of the proposed solution is a machine learning model that can learn to distinguish spammers from non-spammers. The authors chose a Support Vector Machine (SVM) classifier, which is a well-established and effective algorithm for high-dimensional data classification. They used the Radial Basis Function (RBF) kernel, which is a popular choice for non-linear classification problems.
  + The authors evaluated the SVM classifier using a confusion matrix, which shows how many spammers and non-spammers were correctly or incorrectly classified. The results show that the SVM classifier is highly effective, with an accuracy of 99.1% for spammers and 99.9% for non-spammers. This means that only a very small fraction of instances were misclassified.

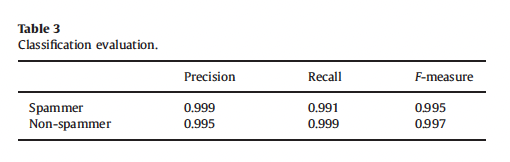


Figure 4.6.3: Table depicting results [1 (ZhipengZeng. et al 2015)]

SybilBelief: A Semi-Supervised Learning Approach for Structure-Based Sybil Detection [43]

* Paper [44] addresses Sybil attacks in social networks, proposing SybilBelief, a semi-supervised learning framework for Sybil detection. The framework incorporates known benign and Sybil labels and utilizes Markov Random Fields for propagation and ranking. It proves resilient to noise and significantly outperforms previous methods.
  + The paper introduces a novel probabilistic local rule to compute reputation scores based on the labels of neighboring nodes. This rule is integral to the SybilBelief framework.
  + The authors clarify how SybilBelief leverages known labels to perform both Sybil classification and ranking, as well as the use of boosting when only labeled benign nodes are available.The paper describes the utilization of Loopy Belief Propagation (LBP) as the inference algorithm for calculating posterior distributions. LBP involves message passing between neighboring nodes.
  + The authors of the paper present a comprehensive comparison between SybilBelief and state-of-the-art Sybil classification and ranking mechanisms. The comparison is based on real-world social network data and demonstrates the superior performance of SybilBelief.
  + The paper evaluates SybilBelief's performance under various conditions and scenarios, considering the impact of factors such as parameter settings, the number of labels, and label noise on its effectiveness. It also compares SybilBelief to state-of-the-art Sybil classification and ranking mechanisms on real-world social network data, demonstrating SybilBelief's superior performance.

###### **Sangeeth Santhosh**

Incorrect classification of social media accounts into suspicious or non-suspicious needs to be avoided at any cost. If this is not taken care of, then the detection process of suspicious activities on social media using data mining and machine learning techniques would be negatively impacted, thus causing many cases of incorrect detection of such activities.

Classifying Suspicious Content on Social Media Networks [33]

* This research paper focuses on addressing the use of sentiment analysis technology to detect suspicious content in tweets generated on Twitter.
* In the method used in the paper, the integrity of honest users is not compromised as only tweets posted by accounts which were later suspended by Twitter are only taken into consideration.
* In the study, it is identified that classification accuracy values of 88.06% and 92.59% respectively are achieved by using Naive Bayes and Random Forest Classifier machine learning algorithms. It is emphasized in the study how crucial it is to classify dubious content on social media, keeping in mind how common misleading information is.
* The following tables show in detail the different values of classification results for Naive Bayes Algorithm and Random Forest Algorithm respectively.

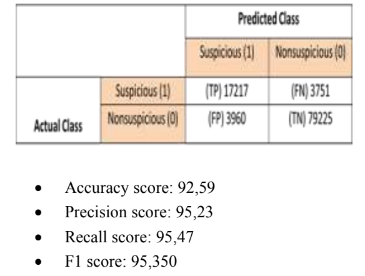
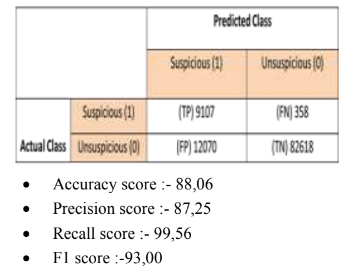


Figure 4.7.1 : Naive Bayes and Random Forest Algorithm Classification Results[33 Ghanem & Habeeb (2021)]

A new approach for the detection and analysis of phishing in social media networks : the case of Twitter [34]

* This research paper focuses on presenting a novel method for identification and analysis of

phishing in social media networks.

* In this paper, the methods used are tested exclusively on real data, but the database that is chosen in each step, blacklist, has only suspicious URLs and websites, thus ensuring that chances of false grouping are minimized to almost a null value. This ensures that integrity of honest users is not compromised.
* The approach that is suggested in the paper consists of mainly three steps: (i) Database search for

questionable URLs, (ii) URL analysis based on machine learning and (iii) Analysis

of user accounts to detect accounts with malicious intent.

* The approach’s application is described, along with characteristics utilized for URL and user account analysis. It reports high accuracy while identifying phishing URLs, especially during use with Random Forest Classifier.
* The results of the different algorithms used are compared effectively in this paper. The comparison graph of the same is shown in the figure below. Clearly, it shows how the Random Forest Classifier is superior as compared to Logistic Regression and Support Vector Machine algorithms.

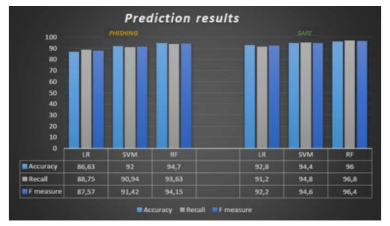


Figure 4.7.2 : Comparison of results [34 (Djaballah et al. 2020)]

* In addition to this, the effectiveness of the tool is also demonstrated in the research paper through testing with actual phishing URLs. This paper aims to extend upon the results identified by expanding it to more social media networks other than Twitter by using more extensive datasets.

Exploring and Detecting Opinion Spam on Social Media [30]

* The research paper focuses on exploration followed by identifying opinion spam on social media websites.
* Opinion spam is when well-organized users post a lot of comments in an attempt to sway public opinion.
* In this paper, the methods used for identifying opinion spam on social media only take into account users who have engaged in opinion spam. Thus, the integrity of legitimate users is not compromised.
* The opinion spam users and legitimate users are differentiated initially based on the frequency of words used by each of these users. The summary of the frequent words used by these two different classes of users are as shown in the table:

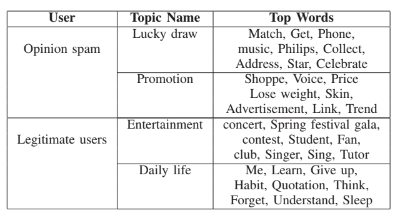


Table 4.7.1 : Word Frequency for Opinion Spam (VS) Legitimate Users [30 Xiao & Qui (2020)]

* The study employs a quantitative approach to investigate the linguistic patterns, behavioral trends and network architectures associated with opinion spam.
* With a 91% f1 score, it suggests a context-based collective classification strategy to identify opinion spam. In conclusion, the authors of this paper gathered a dataset and used a variety of features to show that context and language-based features work well for classifying opinion spam.

Spam Filtering of Bilingual Tweets using Machine Learning [31]

* This research paper focuses on the problem of spam filtering in bi-lingual tweets using machine learning approaches.
* The dataset gathered for this study includes Roman Urdu tweets for the largest cities of Pakistan.
* The effectiveness of classification algorithms like LibSVM, J48, DMNBText, Naive Bayes Multinomial and Liblinear is assessed on this dataset.
* From the results, it is found that Naive Bayes Multinomial and DMNBText outperform the others in terms of accuracy and ROC AUC.
* The performance comparison of each of the algorithms can be summarized as follows, using the table from the research paper:

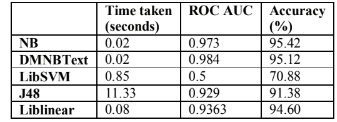


Table 4.7.2 : Comparison of Performance of Different Algorithms [31 Afzal & Memood (2016)]

* The study emphasizes the difficulties that are associated with short-form language and the significance of creating efficient spam detection systems for less widely used languages. The results of the study contribute significantly to enhancing spam filtering in multilingual social media settings.

Ecosystem of Spamming on Twitter: Analysis of Spam Reporters and Spam Reportees [32]

* This research paper investigates the problem of Twitter spamming, with a particular

emphasis on the spammers (reportees) and the users who report them (reporters).

* The paper addresses the problems that Twitter spam poses, groups various kinds of spammers, and explains the characteristics that are employed in the classification process.
* The algorithm proposed in the paper takes a unique approach by focusing on the attributes

of the reporter and reportee. This would help in accurately identifying spam users,

while, also preventing honest users from wrongly getting classified as spam and getting

their Twitter accounts suspended.

* The study classifies reporters and reportees using data mining approaches as Random Forest Classifiers, K-Nearest Neighbors and Decision Trees.
* The following figure shows the implementation pipeline of the proposed approach, with the main steps labeled. It also shows the different data mining techniques that are implemented

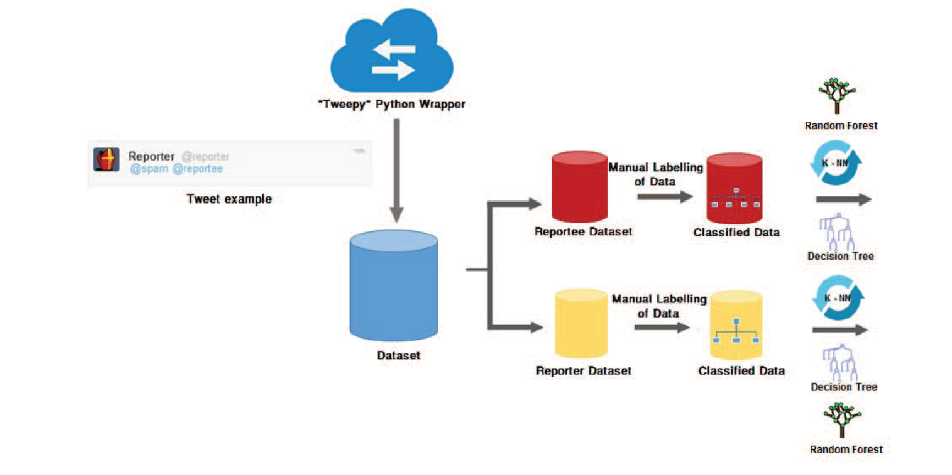


Figure 4.7.3 : Implementation Process of Proposed Approach [32 (Sinha et. al 2016)]

* According to the findings, decision trees offer the highest classification accuracy, followed by Random Forest and K-Nearest Neighbors classifiers.
* Based on the rapport between reporters and reportees, the research indicates that there is room for improvement in the precise identification of spammers.

###### **Yeshwanth Reddy Chennur**

Modern cybersecurity relies heavily on data mining and machine learning. To identify any dangers in the social media data, they use advanced analytics. Machine learning builds detection models, whereas data mining finds patterns. This strategy is essential for protecting online communities and guaranteeing user security.

[A Survey on Security Threats and Defensive Techniques of Machine Learning: A Data Driven View](https://drive.google.com/file/d/1X9JSn0ruiX9H44EjAzVw0RlvPWt-mFfz/view?usp=drive_link)[24].

* This paper reviews the major security threats and defenses for machine learning. Machine learning models can be vulnerable to different types of attacks that mess with integrity, availability, and privacy.
* The authors break down the threats during training and during testing/inference. In training, poisoning attacks sneak in bad data to throw off the model accuracy. Defenses involve scrubbing bad data and making the learning algorithms sturdier. During testing/inference, common threats are evasion to cause misclassification and impersonation to fake being someone else.
* Inversion attacks take advantage of system APIs to steal private training data. Defenses at this stage focus on toughening up models through things like adversarial training, defensive distillation, and smoothing outputs.

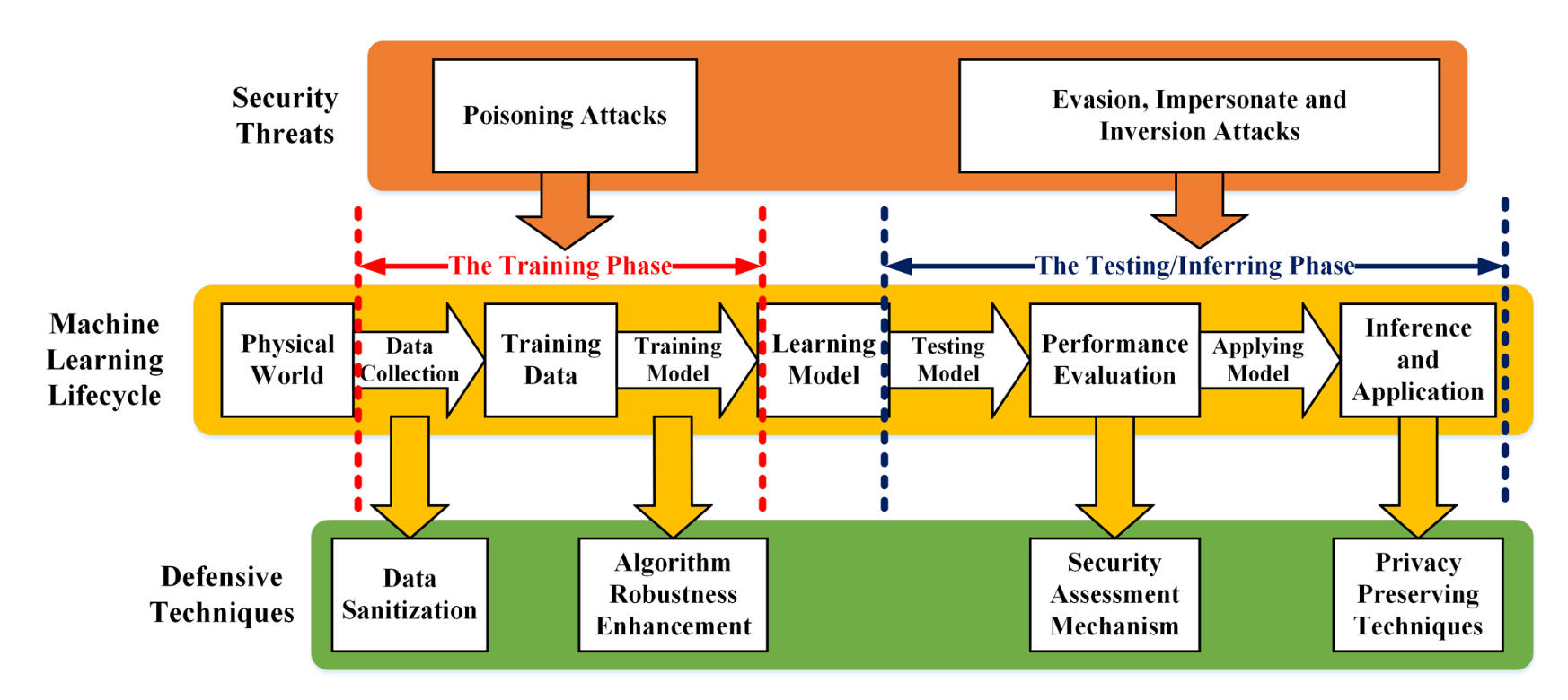


Figure 4.8.1: Illustration of defensive techniques of machine learning [24 (Liu et al. 2018)]

* The authors group defensive techniques into security evaluations, safeguards during training, safeguards during testing/inference, and data security/privacy protections like differential privacy and homomorphic encryption.
* They suggest a proactive threat approach where the system operator simulates attacks to find holes before release.
* The survey spots several key trends in this field: 1) new threat models keep emerging, especially for deep learning; 2) evaluating machine learning security is growing but lacks standards; 3) data privacy techniques remain inefficient for large systems; 4) securing deep learning is still an open issue; and 5) balancing security, accuracy, and overhead during design needs more attention.
* In short, the paper neatly maps out security dangers and defenses across the machine learning pipeline. Key takeaways are that adversarial threats are rising, particularly for deep learning, while blocks remain challenging. The survey serves as a handy reference for future work in this area.

[A Trusted Social Network Using Hypothetical Mathematical Model and Decision- Based Scheme](https://drive.google.com/file/d/15-gauHeasPzFDcXhp6a1OyeXj3vuIdrT/view?usp=drive_link) [23].

* Online Social Networks (OSNs) have witnessed rapid growth in both professional and personal domains. Trust plays a pivotal role in ensuring reliable communication within these diverse user communities.
* However, existing trust computation methods, particularly those based on machine learning, exhibit limitations such as a lack of context-awareness, sparse trust data, and dynamic trust values.
* This paper addresses these shortcomings and aims to establish a robust Trusted Social Network Framework utilizing hypothetical mathematical modeling and decision-making schemes to facilitate secure device-to-device communication through the evaluation of individual node trust.
* The proposed approach involves the categorization of devices into either trusted or malicious entities, predicated on metrics including energy consumption, employing a hypothetical mathematical model.

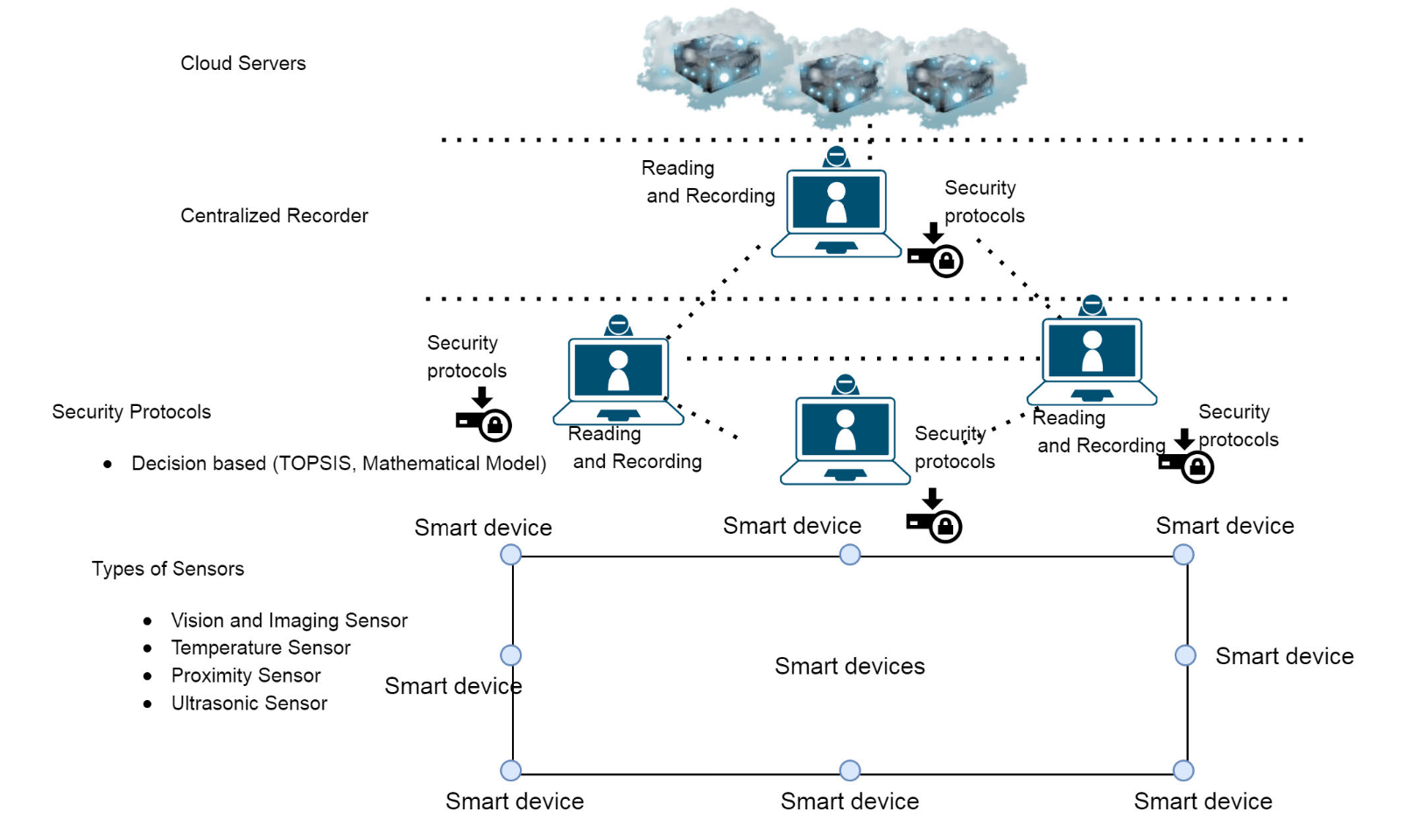


Figure 4.8.2: A trusted social network using hypothetical decision making model [23 (Rathee et al. 2020)]

* Four hypothetical scenarios are defined, contingent on nodes demonstrating authenticity. Error rates are computed for both falsely authenticating malicious nodes and failing to authenticate legitimate ones.
* The integration of Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) expedites actions conducted by the Centralized Authority (CA) overseeing trust. Metrics are subjected to normalization, weighting, and comparison against ideal/non-ideal solutions to effectively categorize nodes.
* This integration of the hypothetical model with TOPSIS facilitates direct trust computation and expedites decision-making processes for enhanced security.
* Simulations conducted on a cohort of 900 Internet of Things (IoT) devices demonstrate that the proposed methodology achieves an approximate accuracy of 88% in identifying malicious nodes, surpassing existing methodologies.
* It demonstrates superior effectiveness in mitigating Distributed Denial of Service (DDoS) attacks and countering data falsification when compared to conventional techniques. Furthermore, it leads to a higher rate of processing legitimate node requests.
* Key outcomes encompass the realization of direct trust computation devoid of reliance on transitive paths, accelerated response times through the integration of decision-making techniques, and heightened precision in detecting malicious nodes. These advancements collectively bolster security within social networking environments.

[Hybrid Deep-Learning-Based Anomaly Detection Scheme for Suspicious Flow Detection in SDN: A Social Multimedia Perspective](https://drive.google.com/file/d/1J819fREISmGoL306TgUEZ6VMFeh9pGU6/view?usp=drive_link) [18].

* The surge in social media usage has led to a notable rise in network security threats. There is an immediate requirement for swift anomaly detection combined with seamless end-to-end delivery of multimedia content within social networks.
* While Software-defined networking (SDN) plays a pivotal role, it faces constraints in runtime security and energy-conscious networking.
* This paper strives to augment SDN dependability through the proposition of a hybrid deep learning-based anomaly detection system and an SDN-supported multi-objective flow routing strategy tailored for social multimedia.
* The anomaly detection module employs an ensemble of Restricted Boltzmann Machine (RBM) with dropout functionality for dimensionality reduction, complemented by a gradient descent-based Support Vector Machine (SVM) with mixed kernel for classification.

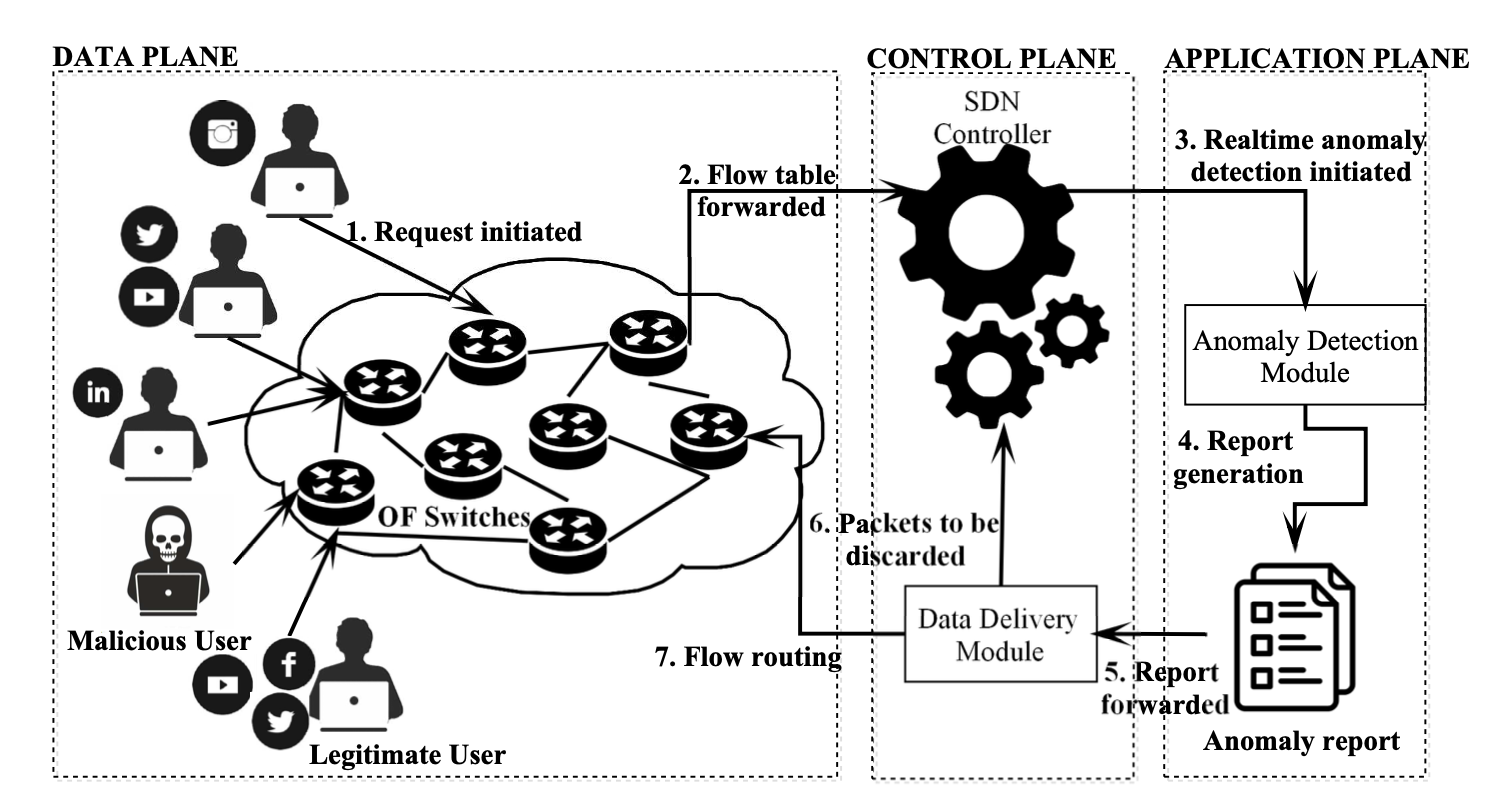


Figure 4.8.3: System model of the proposed SDN-based anomaly detection framework [18 (Garg et al. 2019)]

* The RBM-SVM model conducts real-time anomaly detection. For end-to-end delivery, a multi-objective flow routing (MoFR) scheme is introduced, focusing on minimizing latency and energy consumption while maximizing bandwidth. The MoFR scheme operates on the SDN controller to ensure optimal routing.
* Simulations conducted on both real-time and standard datasets demonstrate that the proposed model achieves an approximate detection rate of 99%, surpassing existing methodologies.
* It adeptly mitigates Distributed Denial of Service (DDoS) attacks and effectively counters data falsification. The MoFR scheme guarantees optimal latency, proficient bandwidth utilization, and prudent energy consumption compared to conventional networks.
* Key outcomes encompass real-time anomaly detection with elevated accuracy and the delivery of high-quality end-to-end multimedia content.
* This integrated framework markedly elevates the security and dependability of SDN for social multimedia.

[Towards automated real-time detection of misinformation on Twitter](https://drive.google.com/file/d/12Lm3n-1HkhbutIC-yg4k6kkTWQTXhiwn/view?usp=drive_link)[25].

* The objective of this research is to address the proliferation of misinformation and rumors on Twitter. Given Twitter's emergence as a major news source, its lack of moderation renders it vulnerable to falsehoods during events like emergencies.
* Twitter rumors can engender widespread panic, instability, and misallocation of government resources. Detecting misinformation rapidly is critical to distinguish credible news from unfounded rumors.
* The authors put forth an automated technique to identify trending Twitter rumors by contrasting tweets from verified news entities versus general unverified users.

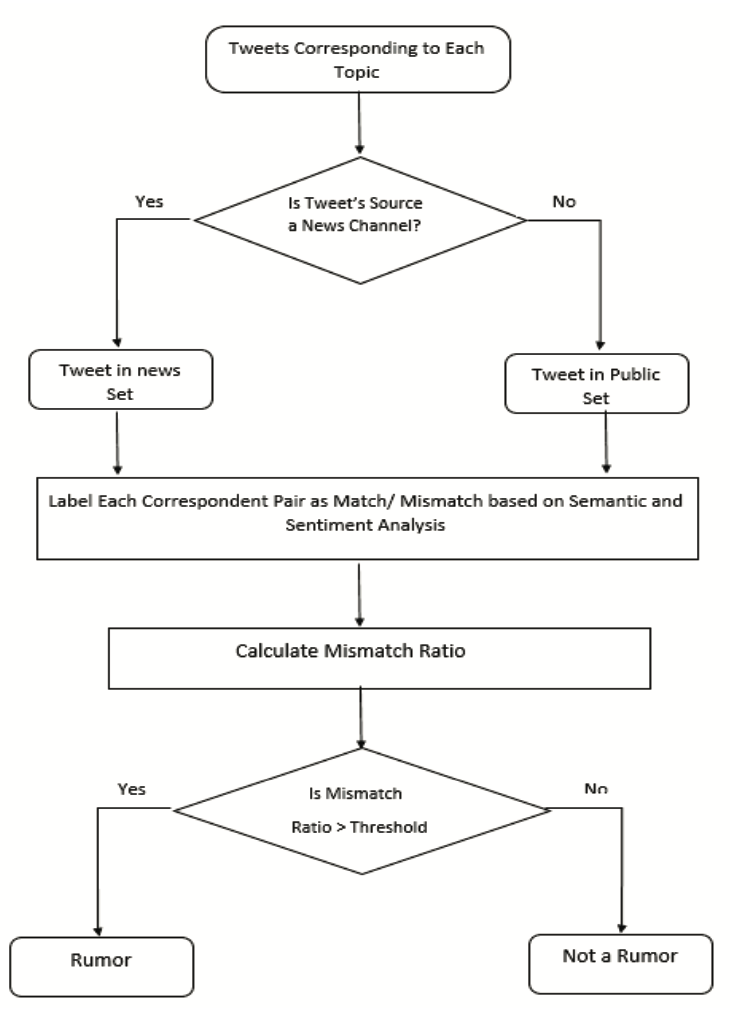


Figure 4.8.4: Rumor detection flowchart [25 (Jain et al. 2016)]

* The proposed methodology collects trending tweets and detects subtopics utilizing hashtag graphs and community detection algorithms. Tweets are partitioned into news tweets from verified accounts and public tweets from unverified users.
* News and public tweets are compared via semantic analysis to ascertain contextual similarity and sentiment analysis to detect polarity divergence.
* High semantic congruity yet opposite sentiment orientations imply a potential rumor. The degree of mismatch between news and public tweets determines if a topic is deemed a rumor.
* The approach is evaluated on real Twitter data such as "Kerala House beef" and "Digital India Facebook" rumors.
* A prototype application visualizes detected rumors with activity timelines, maps, and tweet feeds.
* The approach successfully identified real-world rumors like doctored images of the Paris attacks. On sample topics, it demonstrated 60-77% accuracy in predicting actual rumor labels.
* The system correctly classified factual topics when news and public sentiment converged. The prototype interface enabled interactive rumor monitoring with user feedback.
* The technique shows promise for an automated system to detect trending Twitter misinformation without manual inspection.

[Reveal: Online Fake Job Advert Detection Application using Machine Learning](https://drive.google.com/file/d/1-j6wU2avjIC1IMK1-2n-QTGPmOxD8sIQ/view?usp=drive_link) [35].

* The core goal of this research was to create a machine learning web application named Reveal that can accurately identify fraudulent online job postings.
* Recently, online recruitment scams have increased dramatically, with scammers posting fabricated job ads to steal personal details and money from applicants.
* Many job seekers, especially young graduates, are falling prey by applying for such fake openings. The authors set out to build Reveal to recognize these sham job ads so users can avoid being duped and apply solely to legitimate vacancies.
* First, the researchers accumulated a dataset of verified, real job postings from various sources like company APIs and employment portals. This became the ground truth for comparison. Reveal takes a job URL entered by the user as input.
* It extracts the job information from the URL using web scraping techniques and structures the unorganized HTML content. Natural language processing methods process the text. Key features like title, employer, location, salary are pulled from the extracted data.
* These are stacked against the authentic dataset using machine learning models including random forest, support vector machines, decision trees, linear regression, etc.
* The models are trained on the dataset. If the details match, Reveal categorizes the job posting as valid. If not, it flags it as fraudulent.
* The core contribution is the creation of Reveal, which can effectively identify sham online job postings by contrasting the specifics against a verified database. Extensive testing shows high accuracy, precision and recall in pinpointing fake recruitments.
* A survey revealed most respondents have come across fake ads and Reveal could help avoid applying for such phony openings.

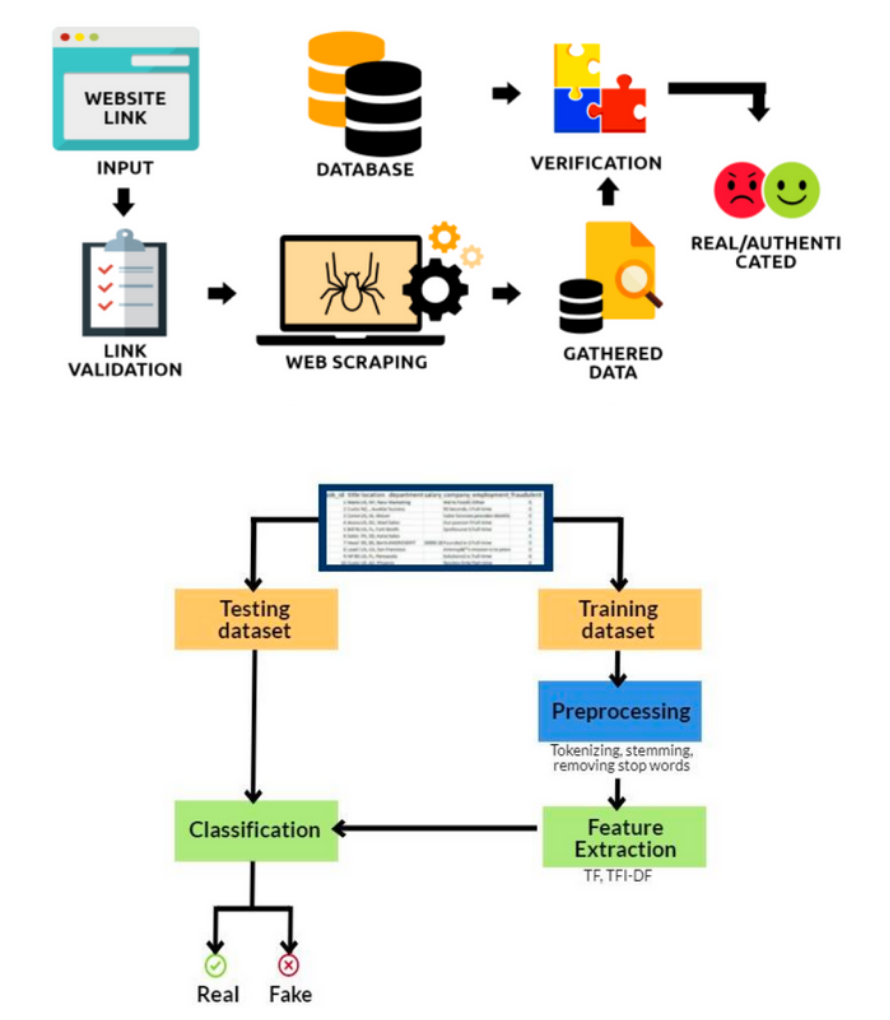


Figure 4.8.5: Classification Flow Chart [35 (C et al. 2022)]

* In essence, the research introduced an AI-powered solution called Reveal to tackle the expanding problem of online job scams. It combines techniques including web scraping, NLP and machine learning to cross-verify job ads against authentic data to catch fraudulent posts.
* This safeguards job seekers by stopping them from getting tricked by the many fake recruitments now found online.

##### **CONCLUSIONS AND RECOMMENDATIONS**

Before we detect and mitigate different types of suspicious activities that happen on social media using machine learning, it is important that we learn what different types of suspicious activities could potentially happen on social media.

Paper [4]:This paper examines different types of suspicious activities on social media: social spam, social Sybils, and link farming. These activities can have harmful effects, such as spreading misinformation and undermining trust. Detecting and addressing them is essential for maintaining the integrity and security of online social environments. This paper provides valuable insights into these different types of suspicious activities, helping researchers and practitioners understand the diverse threats posed by social media and develop strategies to detect and mitigate them.

Paper [2]: Focusing on Twitter and spam type of suspicious activity, this paper addresses the issue of spam detection and the negative impact of spam on the platform. It presents a machine learning framework for detecting spam accounts on Twitter using a variety of features, including content-based, user-based, and graph-based attributes. The paper highlights the importance of feature selection and the challenges in spam detection, emphasizing the evolving tactics of spammers. The findings contribute to the development of effective spam detection models and the protection of users from spam-related threats.

Paper[3]: This paper addresses the critical issue of identifying and mitigating malicious profiles on social media platforms, including fake profiles, spam accounts, bot accounts, and clone accounts. By collecting data from platforms like Twitter, Facebook, and LinkedIn, it highlights the variety of malicious profiles that can harm individuals and online communities. The research employs machine learning algorithms to classify profiles as malicious or genuine and reports results within the range of 80-95% accuracy. This work is crucial in enhancing the security and trustworthiness of social media platforms.

Paper [1]:This paper focuses on detecting spam accounts on a different social media platform Sina Weibo. To detect spam accounts on Sina Weibo, a Chinese social network, this paper proposes a comprehensive approach. The authors collected Sina Weibo data and analyzed both content-based and user-based features to distinguish spammers from non-spammers. They then used a Support Vector Machine (SVM) classifier to detect spam accounts with high accuracy. By analyzing both the content and characteristics of users, the paper offers valuable insights into distinguishing spammers from non-spammers. The SVM classifier's impressive accuracy demonstrates the effectiveness of the proposed solution in identifying and countering spam activities on the platform.

Paper[43]: This paper addresses Sybil attacks in social networks which is another type of suspicious activity, introducing the SybilBelief framework for detection. It incorporates known benign and Sybil labels, utilizing Markov Random Fields and Loopy Belief Propagation for propagation and ranking. The paper demonstrates the framework's resilience to noise and superior performance in comparison to existing methods. By focusing on Sybil attacks, the paper contributes to the understanding and detection of fake or malicious accounts that can deceive and manipulate online communities. The framework presented in the paper offers a promising solution for detecting and mitigating such threats.

Paper [10]: This study concludes with a novel method of identifying malicious tweets using advanced NLP techniques and the notion of a "divergence point." It is admirable that tweets with and without URLs were separated, and that data was gathered via Twitter's API. It is advised to use a bigger and more varied dataset to further validate the model's efficacy.

Paper[12]: The difficulties in interpreting Arabic tweets are discussed in this research, which also offers a methodical way to categorize and preprocess data in order to identify suspicious messages. Consider automating the labeling process to increase efficiency. Moreover, broadening the scope of the research to encompass additional languages and dialects will augment its relevance.

Paper [13]: In summary, this paper's integration of NLP and LSB approaches offers a reliable method for spotting suspicious activity on social networks. Beyond the first testing, it is advised to investigate the approach's scalability and practical application in a wider setting.

Paper [11]: In this paper, a thorough method for spotting questionable user groups on social networking sites is presented. It's great that different NLP and keyword-based systems have been included. To achieve even better outcomes, future research might concentrate on improving topic detection accuracy.

Paper [14]: This study's result emphasizes the value of data preparation and the potency of VADER for sentiment analysis in posts on social media. Increasing the size of the dataset and adding more sophisticated sentiment analysis techniques would improve the effectiveness of the model even more. To increase the predictive power of the system, more data should be gathered.

Paper[43]: This study provides a technique that makes use of NLP and communication pattern analysis to identify malevolent people on social networks. Subsequent investigations may examine its implementation in various social networks and modify the techniques to tackle developing privacy issues in the future.

Paper[42]: To sum up, this work offers a useful strategy for social media tweet analysis that combines NLP with data pretreatment. Longer-term research and the gathering of more data are necessary for future work in order to improve the model's performance under various conditions. The writers ought to think about broadening their scope to include additional social media networks.

Secure Data Collection and Privacy Preservation is essential in order to protect and maintain the user's privacy to maintain ethics, data confidentiality and integrity. The following conclusions were summed up from the in-depth study of the following 7 research papers:

Paper[5]: This paper provides a comprehensive overview of privacy-preserving data mining, outlining the various methods and techniques used to protect sensitive information while allowing for valuable data analysis. It highlights the challenges and considerations in this evolving field and emphasizes the need for further exploration in areas such as mobile data mining, data stream mining, incremental privacy protection data release, and comparative frameworks for evaluating privacy-preserving algorithms.

Paper[6]: The existing PPDM techniques are intensively reviewed and classified based upon their methods that used data modification approaches. The study addresses data provider concerns and explores methods like data modification, cryptographic techniques, and anonymization to ensure privacy throughout the data lifecycle.

Paper[7]: Results from this paper include an in-depth explanation of methods used in machine learning for classifying suspicious content, which are helpful and relevant to the goal of this project. It also provides an analysis between the randomization and secure multiparty computation (SMC) with results. If the SMC approach is used, the process of data mining doesn’t cause breaches of privacy because only selected parties have access to the data.

Paper[8]: The analysis is conducted on a Facebook dataset to assign risk scores to users, with the notion that greater behavioral divergence implies higher risk. This research aims to assess the method's effectiveness. Three data mining approaches are outlined: supervised, semi-supervised, and unsupervised methods. The paper also highlights the increasing threat of cyberattacks on social networking sites, such as Sybil attacks and malware propagation.

Paper[9]: This paper explores sentiment analysis of Indian political tweets using data mining classifiers. Given Twitter's extensive user base and tweet volume, this research is crucial. It introduces a hybrid approach with supervised classifiers, offering insights into Indian users' political sentiments. Rigorous experimentation and data mining techniques establish the k-nearest neighbor classifier's reliable high predictive accuracy of 99.6456% from the analyzed 2,102,52 tweets.

Paper[41]: This paper discusses the 3LP+ privacy-preserving technique for safeguarding multiple sensitive attributes in online social networks (OSNs). It addresses privacy concerns arising from malicious data mining attacks and presents a three-layered approach. Results indicate that 3LP+ can provide better privacy while maintaining higher utility than an existing privacy preserving technique even if an attacker uses a different set of classifiers.

Paper[45]: This research presents a groundbreaking method for extracting and analyzing healthcare-related information from social media, with a particular emphasis on cancer treatments. Data collecting, text processing, behavior analysis, and symptom-medication identification are important elements. To derive insights from user-generated information, data mining is utilized. This strategy might work well for our project.  
  
Thus from an in-depth study of these papers, it is concluded that 3LP+ is the most efficient privacy-preserving technique that provides better privacy than the existing techniques and k-nearest neighbor classifier is the most efficient approach with the highest accuracy than existing predictive approaches.

Data mining is an essential component of detecting suspicious activities in social media. In order to process and analyze textual data, it is crucial to apply data mining techniques to preprocess data.

Paper[16]: This study describes a method for identifying and detecting criminal activity in social media microblog entries by using text mining techniques and the Normalized Compression Distance (NCD). The authors stress the importance of addressing problems with semantic gaps and data sparseness. The report highlights the importance of using semantic analysis to combat cybercrime in social media and advocates for additional advancements to improve execution time and precision in future research.

Paper[15]:The paper presents a novel abusive content detection model. The Multilayer Perceptron (MLP) classifier in the model achieves an accuracy rate of 86%. The study finds that by using this method, abusive text in social media can be more easily identified, highlighting the significance of syntactic and semantic analysis. Future research entails adding sentiment and emotion variables, as well as developing the model to recognize sarcasm across a range of languages.

Paper[37]: The goal of this research study is to apply sentiment analysis and machine learning to fight cyberbullying on social media, especially Twitter. It suggests a three-module preventive structure that combines analysis, decision-making, and user involvement. Cyberbullying was found to be well identified and classified by machine learning models, particularly Random Forest with TFIDF embedding, with an F1 score of 80.8% for identification and 58.4% for classification. The report lays the groundwork for future research in this field and highlights the urgency of combating cyberbullying.

Paper[38]: In order to identify radical and extremist content on Facebook, this research study presents a methodology for content analysis by users. Data extraction, sentiment analysis, extremist content detection, and data storage in HBase are the four primary stages of the framework. The purpose of the proposed framework is to support law enforcement and cybercrime experts in tracking down and looking into digital evidence of radicalism and violence on social media.

Paper[39]: This research study tackles cyberbullying on social media and suggests a deep learning and natural language processing method. It entails gathering data from diverse sources, preprocessing the data, and analyzing the results using a recurrent neural network (RNN). The research highlights the significance of precise techniques for identifying cyberbullying and its possible integration with diverse social media networks.

Machine learning algorithms are utilized across multiple research papers to enhance security and information assurance on social media platforms. These algorithms play diverse roles in automating processes, improving classification, and efficiently detecting threats.

In "Cyberthreat Detection from Twitter using Deep Neural Networks" [22], deep neural networks, including Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) networks, are employed to classify security-related content on Twitter. These algorithms significantly improve the precision of threat intelligence gathering. They also established advanced content classification methodologies.

For Real-Time Threat Alerting the research in "An effective security alert mechanism for real-time phishing tweet detection on Twitter" [21] utilizes the Random Forest machine learning technique for real-time security alerting. This algorithm efficiently distinguishes phishing tweets from legitimate ones, ensuring rapid identification of potential threats in real-world Twitter data.

For Event Classification, "Weakly Supervised Extraction of Computer Security Events from Twitter" [20] explores machine learning algorithms, including One-Class Support Vector Machines (SVMs) and Expectation Regularization, to classify security content on Twitter. The Expectation Regularization technique improves precision and recall in classifying security-related events.

For Fake Account Detection, In "A feature-based approach to detect fake profiles in Twitter" [36], three machine learning algorithms, including Logistic Regression, Support Vector Machines (SVM), and Random Forest, are evaluated for detecting fake Twitter accounts. Random Forest demonstrates superior performance, contributing to the accuracy of fake account detection.

For Alert generation, CyberTwitter uses the SVCE tool to extract cybersecurity data from Twitter [19], filtering for relevance. It employs SWRL rules and RDF triples to create structured knowledge. Personalized threat alerts are generated based on real-time knowledge and user profiles. The system proves effective in extracting timely cybersecurity insights from Twitter data.

In the paper [18], In order to identify suspicious flows in Software-Defined Networking (SDN), especially in social multimedia scenarios, the paper proposes a hybrid deep learning-based scheme. It shows enhanced performance in anomaly detection, providing a viable way to improve security in SDN environments. This study offers important new perspectives on social multimedia application security.

Finally, machine learning algorithms are pivotal in streamlining security processes, improving content classification, and enhancing crime prediction on social media platforms. They provide essential tools for automating tasks, ensuring real-time threat detection, and contributing to the advancement of security and information assurance.

An integral part of the detection process of suspicious activities on social media using the techniques of data mining and machine learning is to ensure that no honest users are incorrectly classified as spam.

Paper [33]: In this research paper titled, ‘Classifying Suspicious Content on Social Media Networks’, the main focus of the authors is to create a model to generate opinions on suspicious and non suspicious measures for tweets on Twitter along with its corresponding classification into these categories using Sentiment Analysis and Machine Learning techniques. In this paper, correct classification of users into spam and non-spam is ensured by taking into consideration only those tweets posted by accounts that were later on suspended by Twitter. This, in turn, ensures that integrity of honest users is not questioned.

Paper [34]: In this research paper titled, ‘A new approach for the detection and analysis of phishing in social networks : the case of Twitter’, a three step approach is proposed by the authors to detect and analyze phishing in Twitter. The methods used in this paper are tested purely on real data, which means that the chances of false grouping of input is almost nil. This ensures that no honest users are incorrectly labeled as spam.

Paper [30]: In this research paper titled, ‘Exploring and Detecting Opinion Spam on Social Media’, the authors focus on detecting and analyzing opinion spam on social media along with the development of a model to achieve the same. In this paper, the methods used take into account only those users who have engaged in opinion spam, which goes a long way in ensuring that integrity of honest users is not compromised.

Paper [31]: In this research paper titled, ‘Spam Filtering of Bilingual Tweets using Machine Learning’, the authors focus on filtering spam content on bilingual tweets in social media and analyzing classification techniques that are used for the same. In the data preparation step of the approach proposed, chances of incorrect classification of legitimate tweets as spam is considerably reduced, meaning integrity of honest users is not compromised in almost all cases.

Paper [32]: In this research paper titled, ‘Ecosystem of Spamming on Twitter: Analysis of Spam Reporters and Spam Reportees’, the authors focus on analysis of spam activity on Twitter along with the corresponding reporters and reportees responsible for the spam. The algorithm that the paper proposes uses the attributes of the reporters and reportees, which help in accurate identification of spam users and prevention of honest users from getting incorrectly classified as spam.

In the paper [24], Based on data-driven insights, the research provides a thorough analysis of security threats and solutions in the machine learning space. For academics and professionals working in this rapidly developing field of study, it is an invaluable resource.

When deploying machine learning-based algorithms and models into social media security problems, it is vital that our models perform accurately and efficiently to classify and detect as many attacks as possible.

In paper [28], the performance of the KNN, Naive Bayes and decision tree algorithms are compared in a mobile application environment against malware-based attacks. The results of the study concluded that Naive Bayes performed the best with an accuracy of 97.27%.

In paper [26], a comparative analysis of machine learning algorithms on a binary classification problem to determine if network traffic is malicious or benign. The algorithms analyzed are XGB, Decision Tree, Random Forest, K-Nearest Neighbor (KNN) classifiers and Multi-Layer Perceptron (MLP) and Quadratic Discriminant Analysis (QDA). The results indicated that the decision tree was the most accurate and efficient in training in this network intrusion detection problem.

In paper [27], a new cloud intrusion detection model SPC was proposed and tested against other machine learning algorithms. The purpose of this research is to measure the performance of the SPC model in detecting and classifying cloud anomalies against complex tree and ensemble models. This study concluded that the SPC model resulted in an accuracy of 81%, which is an improvement of 20% over traditional models.

In paper [37], the cloud environment is analyzed again, however this study focuses on proposing a new concept called “transfer learning”. This method involves the transfer of knowledge from source trained attacks to target attacks. This concept proved to be highly effective when paired with a deep learning model.

In paper [29], SQL-based attacks and XSS cross site scripting are identified as major threats in web intrusion detection and compares the performance of machine learning algorithms in a separate experiment for each attack type. The first experiment is a performance comparison of SQL-based attacks, and concludes that the Adaboost model performed the best due to its high recall score. The second experiment is a text disambiguation process on XSS data, which concluded that the SVM model was the most accurate, however it had drawbacks such as a high false positive rate and a lack of sufficient learning capability.

In the paper [35], An online tool called "Reveal," which uses machine learning to quickly identify fake job postings, is presented in the study. Due to its ability to discern between real and fake job listings, this useful tool protects job seekers from potential scams and deception.

In the paper [23], The suggested framework combines a decision-oriented methodology with an artificial mathematical model to improve trust in social networks. This method gives the network a measurable indicator of trust, allowing it to filter and prioritize material. This is a big step in the direction of building a more dependable and safe online community.

Adaptive learning uses adaptable algorithms that are continuously improved to increase accuracy when used to identify suspicious activity on social media. When it comes to using data mining and machine learning to identify possible security breaches in online social networks, insightful models make use of advanced analytics to obtain a deeper contextual understanding.

In the paper [25], Using computational methods such as data mining and natural language processing (NLP), this research suggests an automatic real-time disinformation detection system on Twitter. It emphasizes how critical it is to identify problems as soon as possible in order to stop the spread of incorrect information on social media.

##### **REFERENCES**

[1] Xianghan Zheng, Zhipeng Zeng, Zheyi Chen, Yuanlong Yu, Chunming Rong, “Detecting spammers on social networks”,Neurocomputing,Volume 159,2015, Pages 27-34,ISSN 0925-2312

[2] S. Gheewala and R. Patel, "Machine Learning Based Twitter Spam Account Detection: A Review," 2018 Second International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2018, pp. 79-84, doi: 10.1109/ICCMC.2018.8487992.

[3] S. Shinde and S. B. Mane, "Malicious Profile Detection on Social Media: A Survey Paper," 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2021, pp. 1-5.

[4] M. Jiang, P. Cui and C. Faloutsos, "Suspicious Behavior Detection: Current Trends and Future Directions," in IEEE Intelligent Systems, vol. 31, no. 1, pp. 31-39

[5] Xinjun Qi, Mingkui Zong, “An Overview of Privacy Preserving Data Mining,” in “International Conference on Environmental Science and Engineering (ICESE)”, Harbin, China, 2011, pp.1341-1347

[6] Mohammed Binjubeir, Abdulghani Ali Ahmed, Mohd Arfian Bin Ismail, Ali Safaa Sadiq, Muhammad Khurram Khan, “Comprehensive Survey on Big Data Privacy Protection,” in “IEEE Access, Vol 12”, Riyadh, Saudi Arabia, 2019, pp.20067-20079

[7] J. Vaidya, C. Clifton, “Privacy-preserving data mining: why, how, and when”, in “IEEE Security & Privacy, Vol 2, Issue 6”, USA, 2004, pp.19-27

[8] M. S. Sudha, K. A. Priya, A. K. Lakshmi, A. Kruthika, D. L. Priya and K. Valarmathi, "Data Mining Approach for Anomaly Detection in Social Network Analysis," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 2018, pp. 1862-1866, doi: 10.1109/ICICCT.2018.8472985.

[9] Anurag P. Jain, Vijay D. Katkar, “Sentiments analysis of Twitter data using data mining,” in “International Conference on Information Processing (ICIP)”, Pune, India, 2015, pp.807-810

[10] S. Gharge and M. K. Chavan, “An integrated approach for malicious tweets detection using NLP,” International Conference on Inventive Communication and Computational Technologies, Mar. 2017, doi: 10.1109/icicct.2017.7975235.

[11] A. Kumar and S. Singh, “Detection of User Cluster with Suspicious Activity in Online Social Networking Sites,” 2013 Second International Conference on Advanced Computing, Networking and Security, Dec. 2013, doi: 10.1109/adcons.2013.17.

[12] M. S. Al-Ghamdi and M. A. Khan, “Intelligent analysis of Arabic tweets for detection of suspicious messages,” Arabian Journal for Science and Engineering, vol. 45, no. 8, pp. 6021–6032, Mar. 2020, doi: 10.1007/s13369-020-04447-0.

[13] M. P. Selvan and R. Selvaraj, “Mining user Message Pattern for Suspicious Behavior on Terrorism using NLP in Social Networks with Single Sign-On,” Indian Journal of Science and Technology, Apr. 2017, doi: 10.17485/ijst/2017/v10i14/111364.

[14] S. A. Mansoori, A. Almansoori, M. Alshamsi, S. A. Salloum, and K. Shaalan, “Suspicious Activity Detection of Twitter and Facebook using Sentimental Analysis,” TEM Journal, pp. 1313–1319, Nov. 2020, doi: 10.18421/tem94-01.

[15] S. Ghosal, A. Jain and D. K. Tayal, "An approach to detect abusive content incorporating Word2Vec and Multilayer Perceptron," 2022 IEEE Bombay Section Signature Conference (IBSSC), Mumbai, India, 2022, pp. 1-5, doi: 10.1109/IBSSC56953.2022.10037274.

[16] S. Alami and O. Elbeqqali, “Cybercrime profiling: Text mining techniques to detect and predict criminal activities in microblog posts,” International Conference on Intelligent Systems: Theories and Applications (SITA), Oct. 2015, doi: 10.1109/sita.2015.7358435.

[17] S. Mitra, T. Tasnim, M. A. R. Islam, N. I. Khan and M. S. Majib, "A Framework to Detect and Prevent Cyberbullying from Social Media by Exploring Machine Learning Algorithms," 2021 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2), Rajshahi, Bangladesh, 2021, pp. 1-4, doi: 10.1109/IC4ME253898.2021.9768450.

[18] S. Garg, K. Kaur, N. Kumar and J. J. P. C. Rodrigues, "Hybrid Deep-Learning-Based Anomaly Detection Scheme for Suspicious Flow Detection in SDN: A Social Multimedia Perspective," in IEEE Transactions on Multimedia, vol. 21, no. 3, pp. 566-578, March 2019, doi: 10.1109/TMM.2019.2893549.

[19] S. Mittal, P. K. Das, V. Mulwad, A. Joshi and T. Finin, "CyberTwitter: Using Twitter to generate alerts for cybersecurity threats and vulnerabilities," 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), San Francisco, CA, USA, 2016, pp. 860-867, doi: 10.1109/ASONAM.2016.7752338.

[20] A. Ritter, E. Wright, W. Casey, and T. Mitchell, “Weakly supervised extraction of computer security events from twitter,” Florence, Italy: International World Wide Web Conferences Steering Committee, 2015, pp. 896–905. doi: https://doi.org/10.1145/2736277.2741083.

[21] Seow Wooi Liew, Nor, Mohd Taufik Abdullah, Razali Yaakob, and Mohd Yunus Sharum, “An effective security alert mechanism for real-time phishing tweet detection on Twitter,” Comput. Secur., vol. 83, pp. 201–207, 2019, doi:https://doi.org/10.1016/J.COSE.2019.02.004.

[22] N. Dionísio, F. Alves, P. M. Ferreira and A. Bessani, "Cyberthreat Detection from Twitter using Deep Neural Networks," 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 2019, pp. 1-8, doi: 10.1109/IJCNN.2019.8852475.

[23] G. Rathee, S. Garg, G. Kaddoum, D. N. K. Jayakody, M. J. Piran and G. Muhammad, "A Trusted Social Network Using Hypothetical Mathematical Model and Decision- Based Scheme," in IEEE Access, vol. 9, pp. 4223-4232, 2021, doi: 10.1109/ACCESS.2020.3048077.

[24] Q. Liu, P. Li, W. Zhao, W. Cai, S. Yu and V. C. M. Leung, "A Survey on Security Threats and Defensive Techniques of Machine Learning: A Data Driven View," in IEEE Access, vol. 6, pp. 12103-12117, 2018, doi: 10.1109/ACCESS.2018.2805680.

[25] S. Jain, V. Sharma and R. Kaushal, "Towards automated real-time detection of misinformation on Twitter," 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Jaipur, India, 2016, pp. 2015-2020, doi: 10.1109/ICACCI.2016.7732347.

[26] E. Osa and O. E. Oghenevbaire, “Comparative Analysis of Machine Learning Models in Computer Network Intrusion Detection”, In 2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development, Lagos, Nigeria, 2022, pp. 1-5, doi: 10.1109/NIGERCON54645.2022.9803175

[27] Z. Chkirbene, R. Camila, A. Erbad, S. Kiranyaz, N. Al-Emadi and M. Hamdi, “Cooperative Machine Learning Techniques for Cloud Intrusion Detection”, In 2021 International Wireless Communications and Mobile Computing, Harbin City, China pp. 837-842, doi: 10.1109/IWCMC51323.2021.9498809

[28] S. Vanjire and M. Lakshmi, “Behavior-Based Malware Detection System Approach For Mobile Security Using Machine Learning, in 2021 International Conference on Artificial Intelligence and Machine Vision, Gandhinagar, India, pp. 1-4, doi: 10.1109/AIMV53313.2021.9671009

[29] C. Chen, J Zhong and W. Chen, “Machine Learning-based web security intrusion detection system”, in 2021 3rd International Academic Exchange Conference on Science and Technology Innovation, Guangzhou, China, pp. 173-177, doi: 10.1109/IAECST54258.2021.9695557

[30] Y. Xiao and J. Qiu, "Exploring and Detecting Opinion Spam on Social Media," 2020 Chinese Automation Congress (CAC), Shanghai, China, 2020, pp. 7570-7573, doi: 10.1109/CAC51589.2020.9327093.

[31] H. Afzal and K. Mehmood, "Spam filtering of bi-lingual tweets using machine learning," 2016 18th International Conference on Advanced Communication Technology (ICACT), PyeongChang, Korea (South), 2016, pp. 710-714, doi: 10.1109/ICACT.2016.7423530.

[32] P. Sinha, O. Maini, G. Malik and R. Kaushal, "Ecosystem of spamming on Twitter: Analysis of spam reporters and spam reportees," 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Jaipur, India, 2016, pp. 1705-1710, doi: 10.1109/ICACCI.2016.7732293.

[33] N. A. Ghanem and H. M. Habeeb, "Classifying Suspicious Content on Social Media Networks," 2021 International Conference on Advanced Computer Applications (ACA), Maysan, Iraq, 2021, pp. 171-175[34] K. A. Djaballah, K. Boukhalfa, Z. Ghalem and O. Boukerma, "A new approach for the detection and analysis of phishing in social networks: the case of Twitter," 2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS), Paris, France, 2020, pp. 1-8

[35] C. Prashanth, D. Chandrasekaran, B. Pandian, K. Duraipandian, T. Chen and M. Sathiyanarayanan, "Reveal: Online Fake Job Advert Detection Application using Machine Learning," 2022 IEEE Delhi Section Conference (DELCON), New Delhi, India, 2022, pp. 1-6, doi: 10.1109/DELCON54057.2022.9752784.

[36] J. Kaubiyal and A. K. Jain, “A feature based approach to detect fake profiles in twitter,” Melbourn, VIC, Australia: Association for Computing Machinery, 2019, pp. 135–139. doi: https://doi.org/10.1145/3361758.3361784. Available: https://doi.org/10.1145/3361758.3361784

[37] G. Sreelatha, A. Vinaya Babu, D. Midhunchakkarvarthy, “Ensure Anomaly-Aware Security Model for Dynamic Cloud Environment using Transfer Learning”, in “2020 5th International Conference on Communication and Electronics Systems (ICCES)”, 2020, pp. 666-670. doi: https://doi.org/10.1109/ICCES48766.2020.9138009

[38] E. Mouhssine and C. Khalid, "Social Big Data Mining Framework for Extremist Content Detection in Social Networks," 2018 International Symposium on Advanced Electrical and Communication Technologies (ISAECT), Rabat, Morocco, 2018, pp. 1-5, doi: 10.1109/ISAECT.2018.8618726.

[39] S. G, D. Chandrasekaran, M. D. Sre and M. Sathiyanarayanan, "Predicting Abnormal User Behaviour Patterns in Social Media Platforms based on Process Mining," 2023 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE), Bengaluru, India, 2023, pp. 204-209, doi: 10.1109/IITCEE57236.2023.10091025.

[40] Khondker Jahid Reza, Md Zahidul Islam and Vladimir Estivill-Castro,”Privacy Preservation of Social Network Users Against Attribute Inference Attacks via Malicious Data Mining”, in Proceedings of the 5th International Conference on Information Systems Security and Privacy (ICISSP 2019), Barcelona, Spain, pp. 412-420, doi: 10.5220/0007390404120420

[41] J. R. Sánchez et al., “Uncovering Cybercrimes in Social Media through Natural Language Processing,” Complexity, vol. 2021, pp. 1–15, Dec. 2021, doi: 10.1155/2021/7955637.

[42] J. Martínez-Romo and L. Araujo, “Detecting malicious tweets in trending topics using a statistical analysis of language,” Expert Systems With Applications, vol. 40, no. 8, pp. 2992–3000, June 2013, doi: 10.1016/j.eswa.2012.12.015.

[43] N. Z. Gong, M. Frank and P. Mittal, "SybilBelief: A Semi-Supervised Learning Approach for Structure-Based Sybil Detection," in IEEE Transactions on Information Forensics and Security, vol. 9, no. 6, pp. 976-987, June 2014, doi: 10.1109/TIFS.2014.2316975.

[44] Sopan Ganpat Sutar, "Intelligent data mining technique of social media for improving health care," in International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 1356-1360, June 2017, doi: 10.1109/ICCONS.2017.8250690.