International Journal of Early Childhood Special Education (INT-JECSE)
DOI: 10.48047/intjecse/V14I6.413 ISSN: 1308-5581 Vol 14, Issue 06 2022

Analysis of Machine Learning Algorithms For Detecting Cyber Crimes On Social Media Twitter

Dr. Ravikant Zirmite
Associate Professor,
MES IMCC College Pune
Pune – Maharashtra, India

Dr.Venugopal Narsingoju
Prof.Yogeshchandra Puranik(*)
Assistant Professor, MCA Department
P.E.S Modern College of Engineering
Pune–Maharashtra, India

Associate Professor,
MES IMCC College Pune
Pune – Maharashtra, India

Abstract:
This research focuses on the identification of cybercrimes occurring on the social media platform Twitter, with the objective of mitigating the rising prevalence and expansion of online security risks. The suggested system employs sentiment analysis methodologies to detect instances of harassment or threats within social media content. Machine learning algorithms, namely Naive Bayes, Logistic Regression, SVM, and CNN are compared to determine their effectiveness in detecting cyberbullying activity. The goal is to achieve the highest accuracy in identifying threats and alerting users to potential criminal activity in real-time. By implementing such a system, users can have a more secure social media experience by staying informed and vigilant about potential risks in both the virtual and real worlds.
Keywords:bitch, jihad, terrorism, defamation, blackmail, radicalization, fool.

Introduction:

In today's digital landscape, social networking platforms have seamlessly integrated into our daily routines, offering extensive opportunities for communication and establishing connections with others. Nonetheless, alongside their advantages, these platforms expose users, particularly the younger generation, to various risks and vulnerabilities, including the perils of cyber-victimization. The proliferation of social media usage on platforms such as Twitter, Facebook, and Instagram have opened new avenues for the dissemination of hate speech and the commission of cybercrimes.

Given the surge in online activities and the prevalence of cyber threats, it has become imperative to institute effective measures for detecting and thwarting such unlawful activities. Internet surveillance plays a pivotal role in identifying potential life-threatening attacks and sifting through suspicious profiles or activities. In this context, our research endeavors to confront the challenge of identifying menacing messages shared by individuals on social media platforms, with a particular emphasis on Twitter.

Twitter, being one of the most widely-used online social networks, enables users to express their opinions via succinct text messages known as "tweets." Due to their concise format, tweets provide a unique avenue for exploring public sentiment. Consequently, we propose harnessing sentiment analysis techniques to identify offensive language and classify it into distinct threat categories.

Our research approach entails leveraging the publish-subscribe messaging pattern offered by social networking platforms, specifically Twitter, to filter out cyber threats. We have devised a filtering system capable of extracting tweets from various regions, spanning from those deemed most perilous to the safest. This geographical analysis empowers us to pinpoint tweets related to criminal activities and perform sentiment analysis, thereby detecting crime-prone regions nearly in real-time.

By implementing this methodology, our aim is to make a meaningful contribution to the field of cybercrime detection and prevention by tapping into the wealth of publicly available data on social media platforms. Through sentiment analysis, we can glean valuable insights into the prevalence and geographical distribution of cyber threats, enabling us to take proactive measures toward fostering a safer online environment.

Literature Review:

Several research papers have explored the use of sentiment analysis and social media data for detecting and preventing cyber threats. In this literature review, we discuss a few notable studies that have investigated similar topics, focusing on the amount of data used, the algorithms employed, and their respective accuracies.

[1]Selma Ayşe Özel, Esra Saraç, Seyran Akdemir, Hülya Aksu et al. (2017) conducted a study on detecting cyberbullying in Turkish texts using machine learning techniques. They gathered data from Instagram and Twitter and discovered that incorporating both words and emoticons as features led to an enhancement in cyberbully detection. The Naïve Bayes Multinomial classifier achieved the top accuracy rate of 84% when feature selection was applied.

[2]Ebraheem Fahad Aljarboua, Marina Bte Md. Din, Asmidar Abu Bakar et al. (2022) conducted a literature review to explore the application of machine learning in cybercrime detection. The study conducted by the researchers established that machine learning models can proficiently discern cybercrime, achieving accuracy levels spanning from 70% to 90%. The primary aim of this investigation was to assess different machine learning algorithms for the automated detection of cybercrime. Eight classifiers, encompassing Logistic Regression, Decision Tree, K-nearest Neighbors, Support Vector Machine, Naive Bayes, Random Forest, eXtreme Gradient Boosting, and Multiple layer perception, underwent scrutiny. The empirical outcomes unveiled that the Multiple layer perception model attained the utmost accuracy, reaching 96% in effectively identifying cybercrime through the utilization of existing data.

[3] MiftaSintaha and Moin Mostakim conducted a study on the use of machine learning algorithms to detect cyberbullying in social media. Utilizing a dataset consisting of 2.5 million tweets, two machine learning models were trained: A Naive Bayes model and a Support Vector Machine (SVM) model. The outcomes demonstrated that SVM outperformed Naive Bayes in terms of accuracy. SVM achieved an impressive accuracy rate of 89.54% in successfully predicting sentiment, while Naive Bayes lagged behind with an accuracy of 73.03%.

[4] Barka Satya, Muhammad Hasan S J, Majid Rahardi, Ferian Fauzi Abdulloh et al. (2022) conducted a study on the sentiment analysis of Sestyc, a social media application popular among Indonesian internet users. By utilizing text data sourced from Google Play Store reviews, the researchers set out to scrutinize user sentiment regarding Sestyc and determine the most proficient sentiment classification algorithm. The study employed three algorithms—Support Vector Machine, Logistic Regression, and Naive Bayes—to assess sentiment. From a pool of 8,000 reviews, 4,719 conveyed positivity, while 3,281 were characterized as negative. Remarkably, the Support Vector Machine algorithm emerged as the frontrunner with the highest accuracy, achieving an impressive 87.81%.

[5] Ankitkumar Patel, Kevin Meehan et al. (2021) investigated the detection of fake news on social media platforms. They conducted a comparative analysis of supervised learning models, including Logistic Regression, Multinomial Naive Bayes (NB), and Support Vector Machine, using both Count Vectorizer and TF-IDF methods on Reddit data. The findings indicated that the combination of Count Vectorizer and the Multinomial NB model exhibited the highest accuracy in discerning fake news. This research serves as a valuable contribution to tackling the task of distinguishing disinformation from genuine content during this era of misinformation. It underscores the significance of Natural Language Processing (NLP) techniques in the fight against fake news.

[6] Abdulkadir Bilen and Ahmet Bedri Özer et al. (2021) conducted a study on cyber-crime detection using machine learning methods. Their examination underscored the efficacy of Support Vector Machine Linear in forecasting cyber-attack methodologies, attaining a remarkable accuracy rate of 95.02%. Meanwhile, Logistic Regression displayed competence in identifying attackers, achieving an accuracy rate of 65.42%. Additionally, the research unveiled a noteworthy association between the education and income levels of victims, suggesting a reduced likelihood of cyber-attacks. These

discoveries yield valuable insights for cybercrime units, facilitating improved cyber-attack detection and prevention efforts.

These studies demonstrate the potential of sentiment analysis in detecting cyber threats and related content on social media platforms. The amount of data used varied across the studies, ranging from thousand to million tweets. Various machine learning algorithms, such as Support Vector Machines, Convolutional Neural Networks, Naive Bayes, Decision Trees, and Random Forest, were utilized for the purpose of sentiment classification. The favorable accuracies achieved serve as evidence of the effectiveness of sentiment analysis techniques in the identification and categorization of cyber threats based on user-generated content.

Proposed System:

Data Extraction:

The proposed system utilizes the Twitter API and Python to extract a large volume of data from Twitter based on specified keywords. Total count of extracted data is 118003 entries.

The process involves:

i. Importing necessary libraries for API access, data manipulation, and environment variables.
ii. Loading API credentials from the environment variables.
iii. Creating an authentication handler and initializing the Tweepy API.
iv. Defining search words and setting the desired number of tweets per query.
v. Extracting tweet texts based on the search words using the API.
vi. Storing the extracted tweet texts in a list.
vii. Convertingthe list into a Pandas DataFrame.
viii. Saving the DataFrame as a CSV file.

This approach efficiently collects real-time tweets matching the keywords, allowing for a significant amount of data extraction.

Preprocessing:The preprocessing step involves cleaning and transforming the collected tweet data to make it suitable for analysis.

Steps performed in the preprocessing stage include:

i. Removing non-English tweets
ii. Converting text to lowercase
iii. Removing URLs, mentions, and hashtags
iv. Removingnon-alphanumericcharacters and special characters
v. Tokenizing the text into individual words
vi. Filtering out words that are not in the English language or not part of selected POS tags
vii. Removing stop words ("a," "an," "the," "is," "are," "in," "on," and soon) and 'rt' (retweet)
viii. Lemmatizing words to their base form

Data Folding: The suggested system applies the data folding method to partition the dataset into training and testing subsets. To accomplish this task, the train-test-split function from the sklearn-model-selection module is utilized. The data is split in an 80:20 ratio, with 80% designated for model training and the remaining 20% allocated for testing and assessing model performance. To ensure data splitting reproducibility, the random-state parameter is fixed at 42.

Automated Training Set Classifier: The code includes the implementation of various classifiers:

i. Naive Bayes (MultinomialNB)
ii. Logistic Regression
iii. Convolutional Neural Network (CNN)
iv. Support Vector Machine (SVM)

Each classifier is trained on the preprocessed tweet data using the corresponding algorithm.

Evaluation: Upon the completion of classifier training, the code proceeds to assess their performance by employing various metrics, including accuracy, confusion matrix, and classification report. The accuracy scores of the trained models, namely Naive Bayes, Logistic Regression, Convolutional Neural Network (CNN), and Support Vector Machine (SVM), are then documented and reported.

Classification of Tweets: The code categorizes tweets into distinct sentiment groups through the application of the trained models. The sentiment categories employed within the code encompass 'Neutral (0)', 'Positive (1)', 'Negative (2)', 'Very Positive (3)', and 'Very Negative (4)'. Each individual

DOI: 10.48047/intjecse/V14I6.413 ISSN: 1308-5581 Vol 14, Issue 06 2022

tweet undergoes classification into one of these categories, a determination made based on its sentiment score or the predicted label it receives.

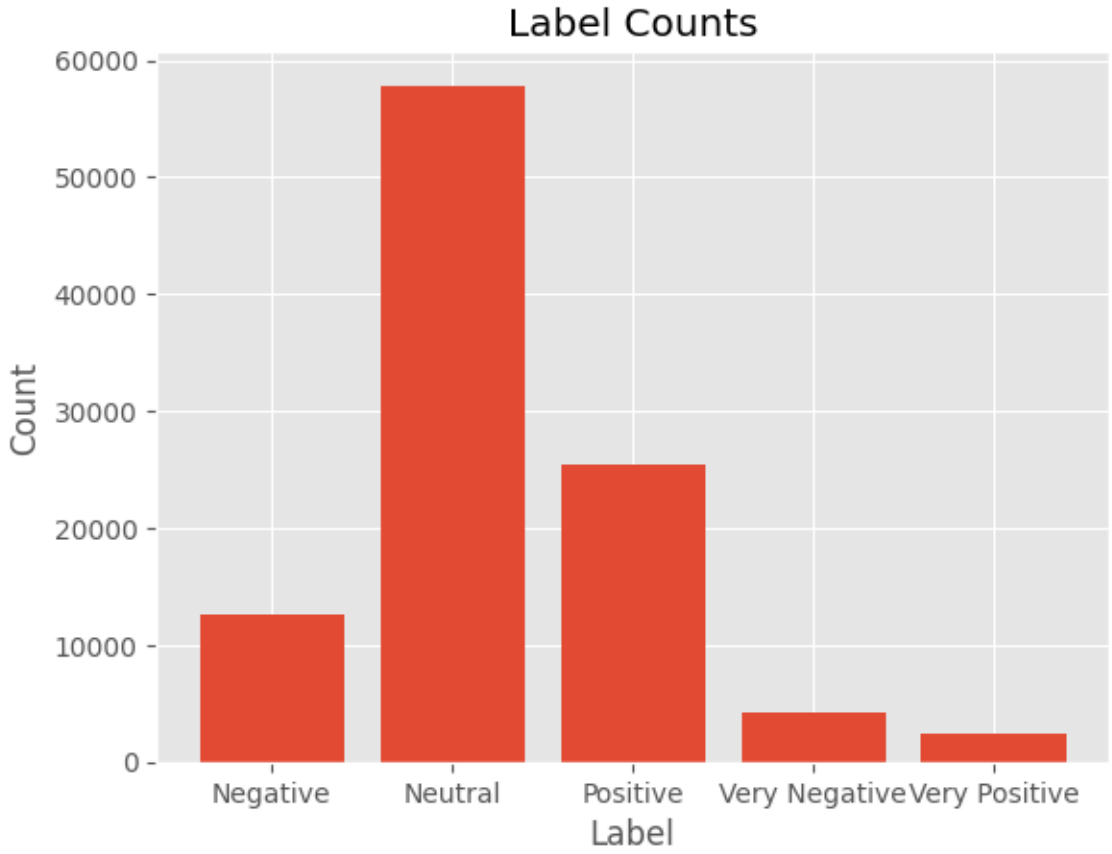

Fig.1 – Sentiment Categories

Feature Extraction:The code employs the TF-IDF (Term Frequency-Inverse Document Frequency) vectorization method to extract features. Utilizing the TfidfVectorizer class from scikit-learn, it transforms the preprocessed tweet text into numerical feature vectors. These extracted features subsequently serve as input for the classifiers.Overall, the proposed system collects data from Twitter, preprocesses the tweets, trains multiple classifiers, evaluates their performance, and classifies new tweets based on sentiment analysis. The provided code encompasses functions for data gathering, data preprocessing, feature extraction, model training, and performance evaluation.

Comparingmachinelearningtechniques

In the section dedicated to the comparison of machine learning approaches, an evaluation is conducted to assess the effectiveness of various algorithms in the detection of cybercrimes on Twitter. Below are the essential aspects concerning the comparison of these machine learning techniques:

DOI: 10.48047/intjecse/V14I6.413 ISSN: 1308-5581 Vol 14, Issue 06 2022

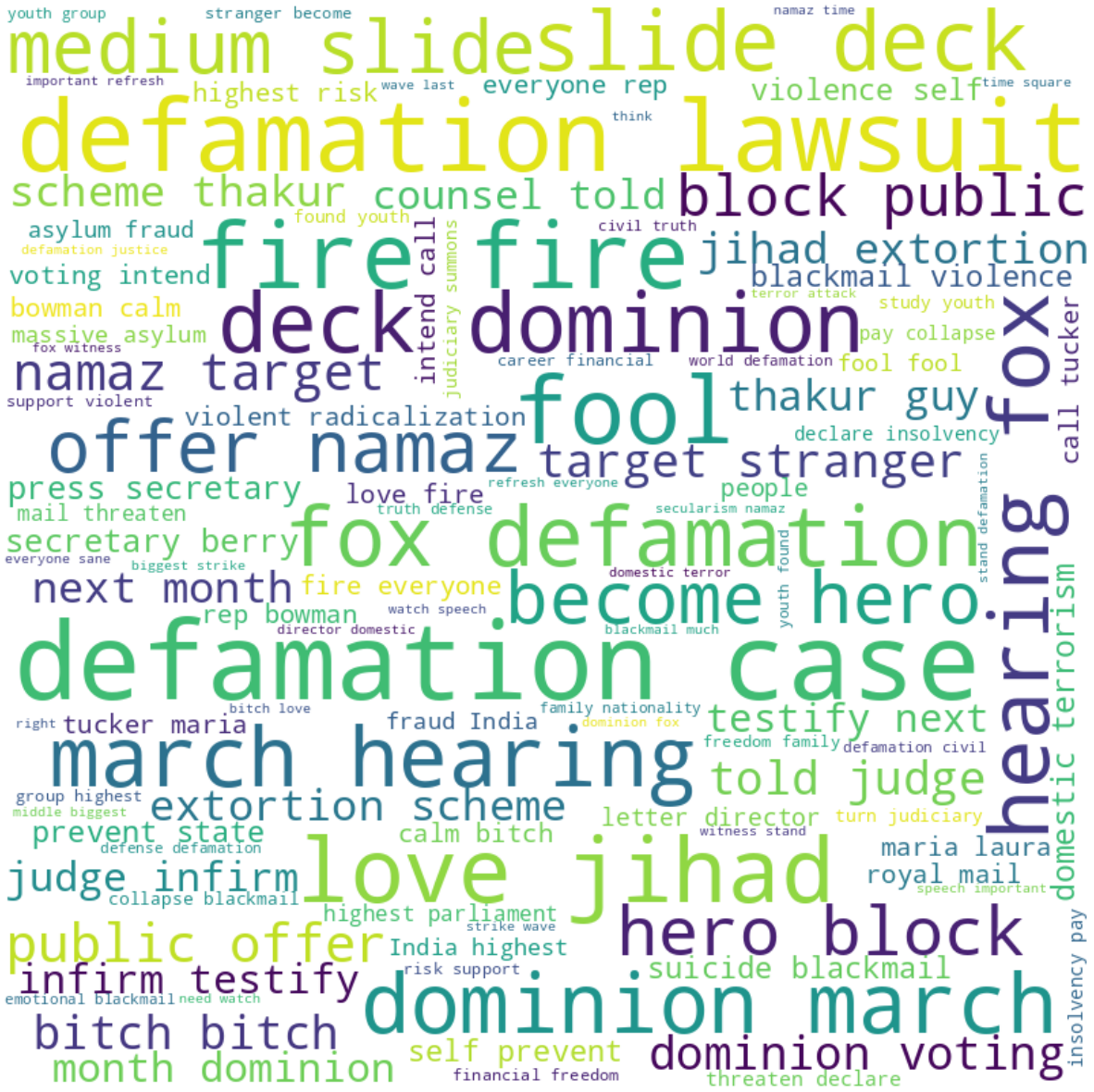

Fig.2 – Most used words

Algorithm	Naïve Bayes	Logistic Regression	CNN	SVM
Accuracy Score	88.32	92.74	94.74	83.92

DOI: 10.48047/intjecse/V14I6.413 ISSN: 1308-5581 Vol 14, Issue 06 2022

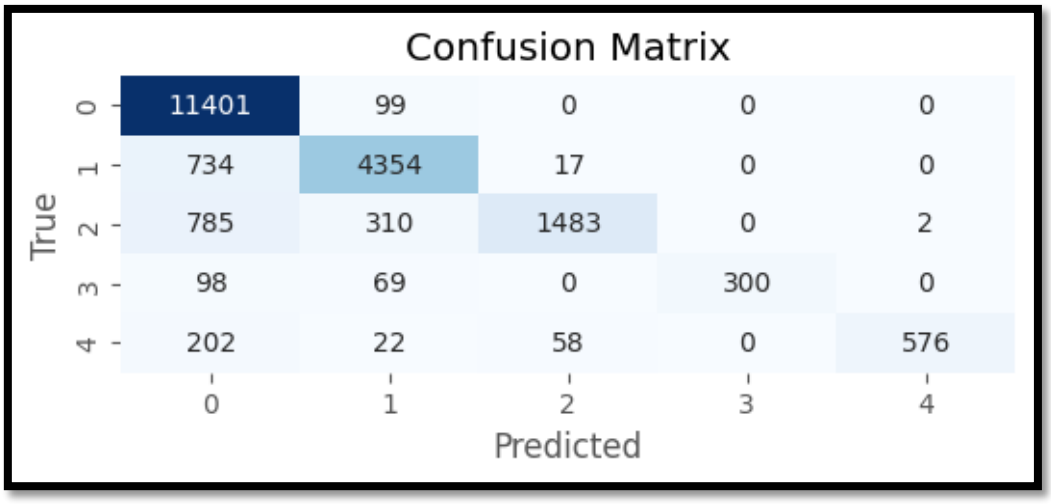

Fig. 3.1 – Confusion Matrix for Naïve Bayes

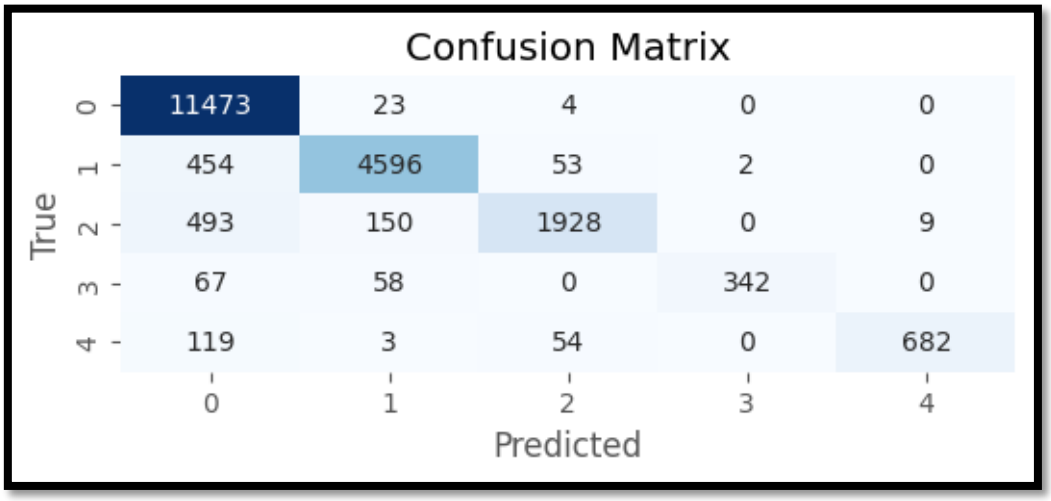

Fig. 3.2 – Confusion Matrix for Logistic Regression

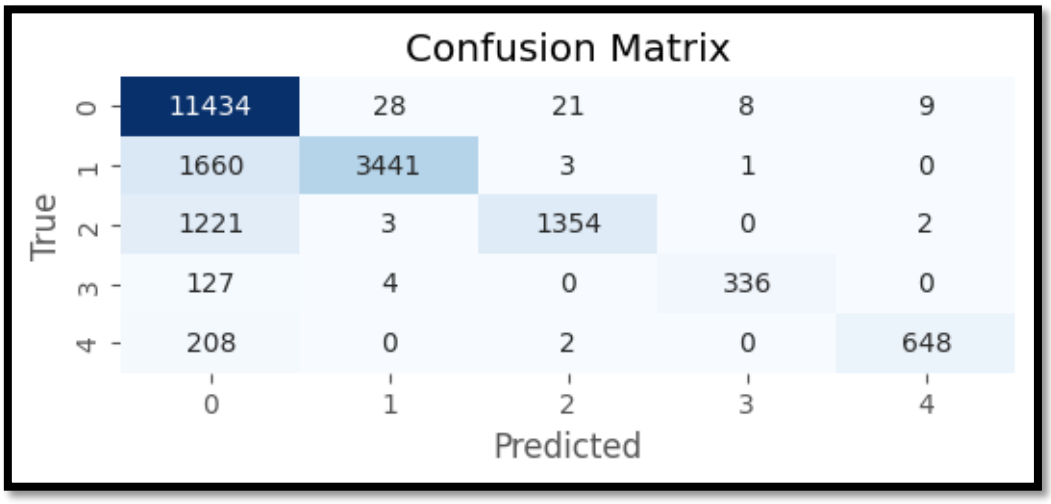

Fig. 3.3 – Confusion Matrix for SVM

Conclusion:

In this research paper, we analyzed different machine learning techniques for detecting cyber-crimes on Twitter. We performed a comparative study of Naive Bayes, Logistic Regression, Convolutional Neural Networks (CNN), and Support Vector Machines (SVM).

Preprocessing and Feature Extraction:

We applied preprocessing techniques to clean the data, remove non-English tweets, URLs, mentions, hashtags, and non-alphanumeric characters.Lemmatization and filtering based on selected parts of speech were performed to reduce noise in the text data.The TF-IDF vectorization technique was used to convert the preprocessed tweets into numerical features for training the models

Performance Comparison:

Naive Bayes achieved an accuracy rate of 88.32%, showcasing commendable performance attributed to its simplicity and efficiency, rendering it well-suited for text classification tasks. Surpassing Naive Bayes, Logistic Regression excelled with an accuracy score of 92.74%, providing valuable insights into the interplay between features and the target variable. CNN, a deep learning technique, attained the highest accuracy score at 94.74%, proficiently capturing intricate patterns and dependencies within the text data, which renders it particularly suitable for classification tasks. SVM secured an accuracy score of 83.92%, underscoring its effectiveness in managing non-linear relationships and high-dimensional data.

Model Selection:

The choice of the most suitable technique depends on factors such as the specific problem, dataset characteristics, and computational requirements.If interpretability and efficiency are crucial, Naive Bayes or Logistic Regression can be considered.For tasks involving complex patterns and hierarchical representations, CNN is a suitable choice.In scenarios involving high-dimensional data and the need for non-linear decision boundaries, SVM can emerge as a viable and effective choice.

Practical Implications:

The discoveries in this research paper have the potential to be employed in the creation of more resilient and precise systems for cybercrime detection on social media platforms such as Twitter. The integration of machine learning techniques can provide valuable assistance to law enforcement agencies and social media platforms in the proficient identification and mitigation of cybercrimes. To summarize, our comparative analysis indicates that CNN achieved the highest accuracy, followed by Logistic Regression and Naive Bayes. The selection of a technique should consider diverse factors, allowing practitioners to choose the most suitable approach based on their specific requirements and the inherent characteristics of the problem they are addressing.

References:

1. Selma Ayşe Özel, Esra Saraç, Seyran Akdemir, Hülya Aksu , “Detection of cyberbullying on social media messages in Turkish”, 2017 International Conference on Computer Science and Engineering (UBMK), Publisher: IEEE.
2. Ebraheem Fahad Aljarboua, Marina Bte Md. Din, Asmidar Abu Bakar, “Cyber-Crime Detection: Experimental Techniques Comparison Analysis”, 2022 International Visualization, Informatics and Technology Conference (IVIT), Publisher: IEEE.
3. Barka Satya, Muhammad Hasan S J, Majid Rahardi, Ferian Fauzi Abdulloh, “Sentiment Analysis of Review Sestyc Using Support Vector Machine, Naive Bayes, and Logistic Regression Algorithm”, 2022 5th International Conference on Information and Communications Technology (ICOIACT), Publisher: IEEE.
4. MiftaSintaha, Moin Mostakim, “An Empirical Study and Analysis of the Machine Learning Algorithms Used in Detecting Cyberbullying in Social Media”, 2018 21st International Conference of Computer and Information Technology (ICCIT), 21-23 December, 2018.
5. Ankitkumar Patel, Kevin Meehan, “Fake News Detection on Reddit Utilizing CountVectorizer and Term Frequency-Inverse Document Frequency with Logistic Regression, MultinominalNB and Support Vector Machine”, 2021 32nd Irish Signals and Systems Conference (ISSC), Publisher: IEEE.
6. Abdulkadir Bilen and Ahmet Bedri Özer, “Cyber-attack method and perpetrator prediction using machine learning algorithms”, http://dx.doi.org/10.7717/peerj-cs.475