Real-Time Social Media Threat Detection Using a Hybrid Semantic Embedding and Multi-Model Cyber Analytics Framework

Nasam Rachana Devi[1], T. Sanath Kumar[2*], Mamidala Swarna[1], Ettaboina Vijay Kumar[1], Godasi Sahasra[1], Arutla Sanath Kumar[1]

[1]UG Student, [2]Assistant Professor, [1,2]Department of Computer Science and Engineering (AI&ML)

[1,2]Vaagdevi Engineering College, Bollikunta, Warangal, 506005, Telangana, India

*Correspondence: T. Sanath Kumar (sunny.554@gmail.com)

Abstract

The rapid growth of social media and digital communication platforms has significantly increased the volume of cybersecurity-related information shared in real time. Platforms such as Twitter and online forums generate massive amounts of unstructured textual data containing valuable insights into cyber threats, vulnerabilities, and ongoing attacks. To overcome these limitations, this study proposes a hybrid NLP-based cyber intelligence framework for real-time threat detection from social media data. The framework begins with preprocessing steps, including tokenization, normalization, and noise removal, to enhance data quality. It then utilizes Sentence-Bidirectional Encoder Representations from Transformers (SBERT) to generate semantic embeddings, enabling a deeper understanding of contextual relationships within textual content. To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) is applied, ensuring balanced datasets for effective training. The system integrates multiple classifiers, including Stochastic Gradient Descent (SGD), Complement Naïve Bayes (CNB), and a hybrid model combining Dense Neural Networks (DNN) with Linear Discriminant Analysis (LDA). This multi-model architecture improves feature representation and classification performance. Experimental results demonstrate enhanced accuracy and reliability, offering a scalable and efficient solution for automated cyber threat intelligence using large-scale social media data.

Keywords: classification task, cyber threats, natural language processing, semantic embeddings, social media, text mining

1.Introduction

With the rapid growth of social media platforms, online communication has become a primary channel for global information exchange [1]. Microblogging platforms enable users to share real-time updates, opinions, and news instantly. Despite these advantages, such platforms have also become mediums for spreading harmful content, including threats, misinformation, and cyber-related activities. The open and accessible nature of social media allows malicious actors to exploit these systems to coordinate attacks or create panic, impacting individuals, organizations, and national security. In recent years, the massive increase in user-generated content has made manual monitoring highly impractical [2]. Millions of posts are generated every minute, making it difficult for authorities to detect potential threats efficiently.

Natural language processing techniques have shown significant promise in analysing social media text and identifying patterns associated with threatening or harmful content [3]. These techniques enable automated systems to interpret textual data, identify contextual meanings, and recognize linguistic patterns that may indicate potential threats. By processing large volumes of online posts, intelligent text analysis systems can assist security agencies and digital platforms in identifying suspicious behavior and responding more effectively to emerging risks. However, the dynamic and informal nature of social media language presents several challenges for threat detection systems [4]. Users often employ slang,

abbreviations, emojis, or coded language, making it difficult for traditional monitoring systems to accurately interpret the true meaning of messages.

Additionally, the rapid evolution of online communication patterns requires analytical systems that can adapt quickly and analyse content efficiently without compromising detection accuracy. To address these challenges, integrated analytical frameworks that combine advanced text analysis with cybersecurity monitoring strategies are increasingly being explored [5]. Such hybrid approaches aim to strengthen the ability of automated systems to detect potential threats in real time by analysing textual signals alongside contextual indicators. These systems can significantly enhance the capability of digital platforms and security agencies to monitor online spaces and respond promptly to potential security risks.

Problem Definition: The rapid growth of social media platforms has resulted in massive volumes of unstructured and fast-moving textual data, making real-time identification of cybersecurity threats increasingly complex. Ambiguous language, noisy content, slang, and evolving adversarial terminology limit the effectiveness of traditional keyword-based or rule-driven detection methods. Additionally, data imbalance and inconsistent reporting of cyber incidents introduce bias into machine learning models, reducing prediction reliability. The dynamic and adaptive nature of cyber threats further complicates automated classification and timely response.

2. Literature Survey

Horvat, et al. [6] introduced a sentiment score aligned with discrete and dimensional emotion models, reliability metrics, and individual word scores using affective datasets Extended ANEW and NRC WordEmotion Association Lexicon. Silvestri, et al. [7] contributed towards an effective threats and vulnerabilities analysis by adopting Machine Learning models, such as the BERT neural language model and XGBoost, to extract updated information from the Natural Language documents largely available on the web, evaluating at the same time the level of the identified threats and vulnerabilities. Saias, et al. [8] examined recent advances in NLP for detecting message-based threats in digital communication. We conducted a systematic review following PRISMA guidelines, to address four research questions.

Merayo, et al. [9] presented a hybrid deep learning model combining convolutional and long short-term memory layers to detect polarity levels in Twitter for the Spanish language. Their model significantly improved the accuracy of existing approaches by up to 20%, achieving accuracies of around 76% for three polarities (positive, negative, neutral) and 91% for two polarities (positive, negative). Mahmud, et al. [10] proposed a hybrid deep learning approach that combines Bidirectional Gated Recurrent Units (Bi-GRUs) and Convolutional Neural Networks (CNNs), referred to as CNN-Bi-GRU, for the accurate identification and classification of smishing attacks. Topcu, et al. [11] analysed data from the Twitter platform and deploy machine learning techniques, such as word categorization, to identify vulnerabilities and counteract zero-day attacks swiftly. TensorFlow was utilized to handle the processing and conversion of raw Twitter data, resulting in significant efficiency improvements.

Hamed, et al. [12] extracted features based on sentiment analysis of news articles and emotion analysis of users' comments regarding this news. These features were fed, along with the content feature of the news, to the proposed bidirectional long short-term memory model to detect fake news. Arora, et al. [13] developed a conversational chatbot, an application that uses artificial intelligence (AI) to communicate, and deploy it on social media sites (e.g., Twitter) for cyber security purposes. Chatbots have the capacity to consume large amounts of information and give an appropriate response in an efficient and timely manner, thus rendering them useful in predicting threats emanating from social media. Raj ,et al. [14] proposedTerm Frequency-Inverse Document Frequency (TF-IDF) demonstrates

consistently high accuracies with traditional machine learning techniques. Global Vectors (GloVe) perform better with neural network models. Bi-GRU and Bi-LSTM worked best amongst the neural networks used. Saeed, et al. [15] investigated how companies can employ CTI to improve their precautionary measures against security breaches. The study follows a systematic review methodology, including selecting primary studies based on specific criteria and quality valuation of the selected data.

Rekha Gangula et al. [16] proposed a sentiment analysis model for Twitter data related to COVID-19. The system performed Natural Language Processing (NLP) preprocessing and classification. The model analyzed public sentiment trends effectively. Rekha Gangula et al. [17] proposed a conceptual framework for understanding machine learning in Artificial Intelligence (AI). The study analyzed various algorithms and their applications. The framework provided insights into AI system design.

Rekha Gangula et al. [18] proposed an analysis of machine learning algorithms in data mining applications. The study evaluated algorithm performance across datasets. The framework improved knowledge discovery processes. Lingala Thirupathi et al. [19] proposed a false news recognition system using machine learning techniques. The framework extracted textual features from news data. The classifier identified misinformation with improved accuracy. Lingala Thirupathi et al. [20] proposed a Twitter sentiment analysis approach using Naive Bayes classification. The system performed feature extraction and polarity classification. The model improved sentiment prediction performance.

3. Proposed Methodology

The research presents a structured framework to analyse cybersecurity-related textual data and detect cyberattack information using machine learning techniques. The system follows a pipeline starting from data ingestion, preprocessing, and transformation of raw text into structured formats using natural language processing. Semantic embedding methods are applied to convert text into numerical vectors capturing contextual relationships. These feature representations enable effective analysis of cybersecurity discussions. Multiple machine learning models are then used to classify and identify cyber threat information as shown in Fig.1.

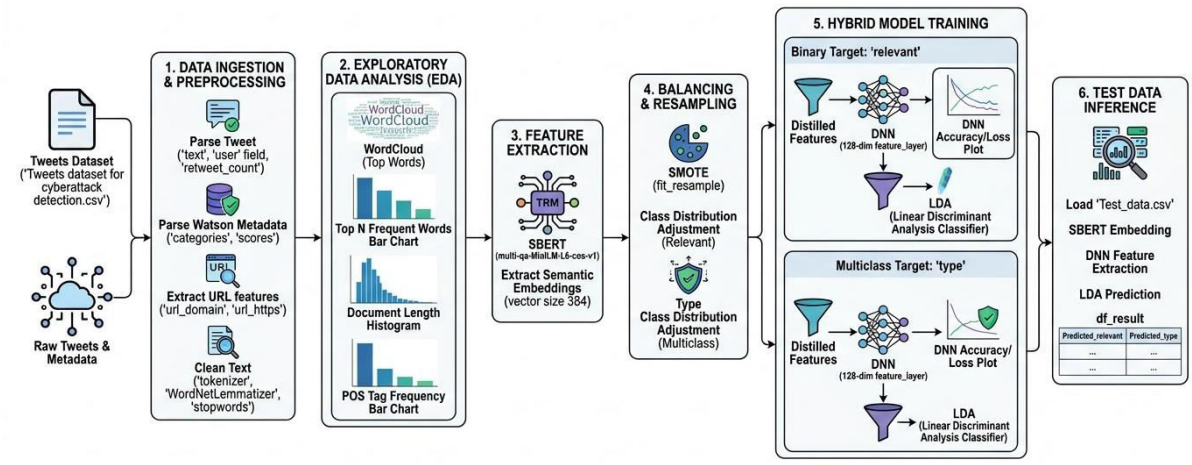


Fig.1: System architecture.

The analytical pipeline integrates baseline classification models and a hybrid deep-learning-based classification approach to improve prediction performance. A Flask-based web application is integrated into the system to provide a user interface that enables dataset upload, prediction generation, and result visualization. The framework also includes evaluation mechanisms that analyse model performance using various statistical metrics and graphical representations. Through this structured analytical

process, the study demonstrates how intelligent machine learning techniques can assist in identifying and categorizing cyberattack-related information from large-scale textual datasets.

User Interface (Web Browser)

- The user interacts with the system through a browser-based interface.
- Users can perform actions such as registration, login, dataset upload, and prediction generation.
- The interface allows users to upload cybersecurity datasets in CSV format.
- All user actions are converted into HTTP requests and sent to the Flask server.

Flask Web Server (app.py)

- The Flask backend receives requests from the web interface and processes them.
- It manages user authentication including registration, login, and session handling.
- The server handles dataset uploads and passes the data to the preprocessing pipeline.
- It loads trained machine learning models and performs prediction tasks.
- The prediction results are returned to the user interface for visualization and download.

Database (MySQL – tweet_db)

- The database stores user information and login credentials.
- It manages records related to user registration and authentication.
- The Flask server interacts with the database for inserting and retrieving user data.
- This ensures secure access to the prediction system.

Dataset Input (Cybersecurity Tweet Dataset – CSV)

- The dataset contains cybersecurity-related textual information collected from online platforms.
- It includes tweet content, user metadata, URLs, and contextual attributes.
- The dataset is uploaded by the user through the web interface.
- This data is passed to the preprocessing module for further analysis.

Data Preprocessing and Cleaning

- The raw dataset is processed to remove noise, irrelevant symbols, and formatting inconsistencies.
- NLP techniques such as tokenization, stop-word removal, and lemmatization are applied.
- Additional features such as user attributes and URL information are extracted.
- The processed data is transformed into structured textual inputs.

Feature Extraction using SBERT

- Semantic embedding techniques are used to convert textual data into numerical feature vectors.
- SBERT generates contextual embeddings that capture semantic relationships in cybersecurity discussions.
- These embeddings represent the textual content in a high-dimensional feature space.
- The generated feature vectors are used as input for machine learning models.

Dataset Balancing using SMOTE

- The dataset may contain class imbalance among different cyberattack categories.
- SMOTE generates synthetic samples for minority classes.
- Balanced datasets improve the learning capability of machine learning models.
- This step reduces classification bias toward majority classes.

Existing Baseline Models (SGD and CNB)

- The extracted features are evaluated using baseline machine learning classifiers:
  - SGD: Performs fast large-scale classification using stochastic gradient optimization.
  - CNB: Performs probabilistic text classification suitable for imbalanced datasets.
- These models provide baseline results for comparison.

Deep Feature Learning using DNN

- The feature vectors are passed through a DNN architecture.
- Multiple dense layers learn complex relationships within the cybersecurity data.
- A hidden feature layer extracts high-level feature representations.
- These deep features capture complex patterns within the dataset.

Final Classification using LDA

- The deep features extracted from the neural network are passed to the LDA classifier.
- LDA performs discriminative classification based on statistical distributions of features.
- It predicts cyberattack type and relevance categories from the dataset.
- This hybrid approach improves classification accuracy.

Prediction Results and Output

- The system generates prediction results based on trained models.
- The predicted cyberattack categories and relevance labels are displayed to the user.
- The results are presented through the web interface.
- Users can also download the prediction results as a CSV file.

3.1 NLP Preprocessing

The NLP preprocessing stage prepares raw cybersecurity textual data for machine learning analysis by converting unstructured information into a structured and meaningful format. The raw dataset often contains noisy text, irrelevant symbols, and inconsistent formatting that cannot be directly processed by analytical models. Therefore, several natural language processing techniques are applied to clean, normalize, and transform the data. This stage ensures that the textual information becomes consistent and suitable for feature extraction and classification processes as shown in Fig. 2.

Data Loading and Initial Parsing: The preprocessing stage begins by loading the dataset containing cybersecurity-related textual information. The system reads the CSV dataset and extracts relevant fields such as tweet content, user information, and associated metadata. This initial step ensures that all required attributes are properly structured and ready for further processing.

Text Cleaning and Normalization: In this stage, the raw text is cleaned to remove unnecessary characters, symbols, and formatting inconsistencies. The text is converted into lowercase to maintain uniformity across all textual entries. Cleaning the text helps reduce noise in the dataset and improves the quality of the information used for analysis.

Tokenization: Tokenization is the process of breaking the cleaned text into smaller units called tokens. Each sentence or document is divided into individual words that can be analyzed independently. This step allows the system to process textual information at the word level and prepare it for further linguistic processing.

Stop Word Removal: Many commonly used words such as articles, conjunctions, and prepositions do not contribute significant meaning to text analysis. These words are known as stop words and are removed during preprocessing. Eliminating such words reduces unnecessary data and helps the models focus on important terms related to cybersecurity discussions.

Lemmatization: Lemmatization converts words into their base or root form while preserving their contextual meaning. For example, words like "attacking" and "attacked" are converted to the root word "attack." This process ensures that similar words are treated as the same feature during analysis.

Feature Preparation: After the text has been cleaned and processed, the relevant textual fields are combined to create a unified representation of the information. Numeric attributes such as user metrics and URL features are also integrated with the processed text. This final representation prepares the dataset for the feature extraction stage where semantic embeddings are generated.

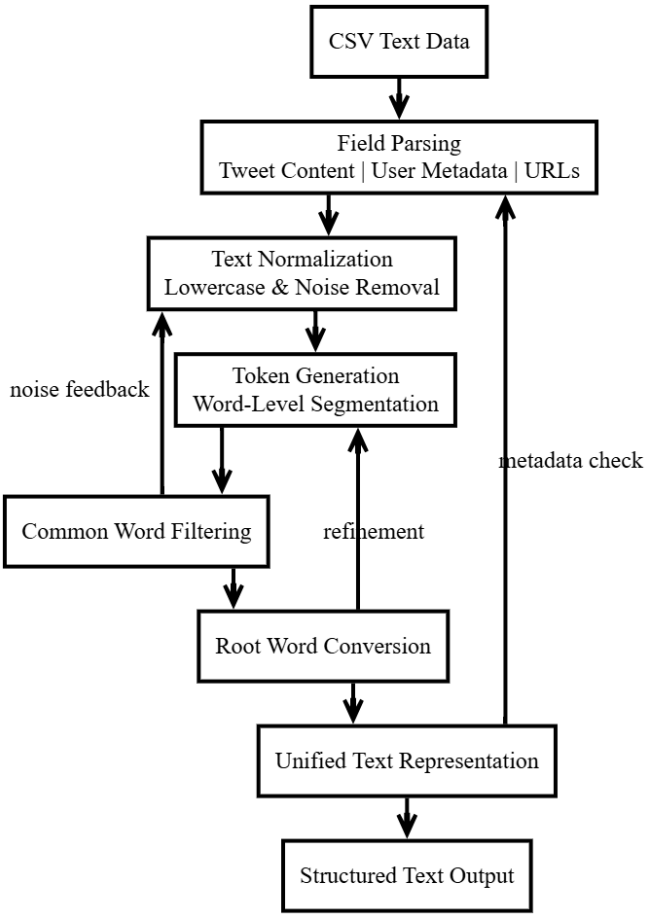
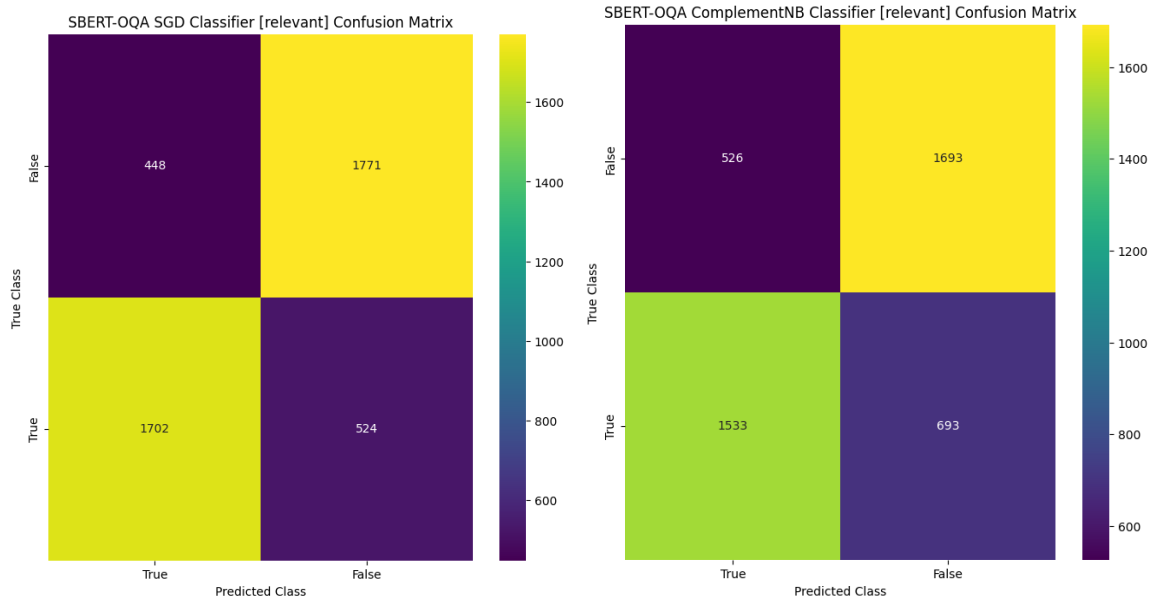


Fig. 2: Internal workflow of NLP preprocessing.

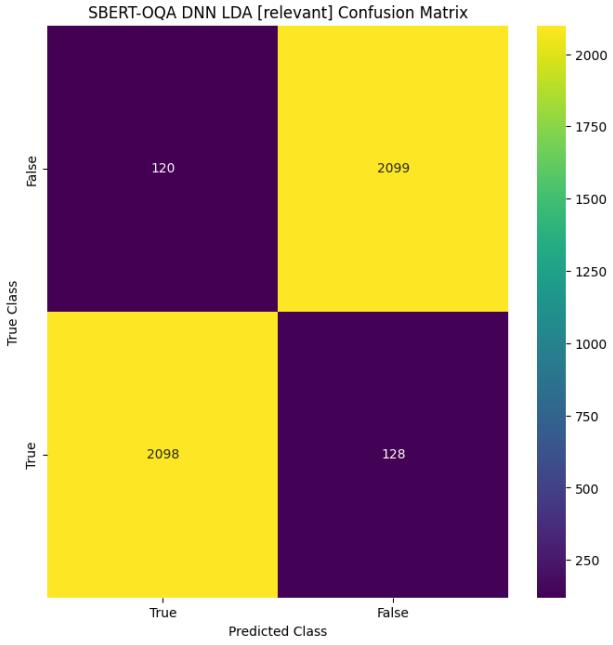
4. Results and Discussion

Fig. 3(a) The SBERT-OQA SGD Classifier's confusion matrix highlights a total of 4,445 instances, with a true positive rate of 1,702 and a true negative rate of 1,771. The relatively low false positive (448) and false negative (524) counts suggest good precision and recall, though there's room for improvement in reducing false negatives for better sensitivity. Fig. 3(b) For the SBERT-OQA CNB Classifier, the matrix totals 4,445 instances, with 1,533 true positives and 1,693 true negatives. The higher false negative count (693) compared to false positives (526) indicates a tendency to under-predict the "true" class, which might affect the model's recall for positive instances. Fig. 3(c) The SBERT-OQA DNN+LDA matrix, also based on 4,445 instances, excels with 2,098 true positives and 2,099 true negatives, alongside minimal false negatives (128) and false positives (120). This suggests a highly accurate model, likely benefiting from the deep neural network and latent Dirichlet allocation (LDA) combination, offering robust classification for the "relevant" column.



(a)

(b)



(c)

Fig. 3: Confusion matrix of target “relevant”. (a) SBERT-OQA SGD Classifier. (b) SBERT-OQA CNB Classifier. (c) SBERT-OQA DNN+LDA

The performance comparison of classification models for the "relevant" column, as outlined in Table 1, reveals distinct differences in effectiveness. The SBERT-OQA SGD Classifier achieves an accuracy of 78.133%, with precision at 78.165%, recall at 78.135%, and an F1-score of 78.127%, indicating a solid and balanced performance. In contrast, the SBERT-OQA CNB Classifier lags with an accuracy of 72.576%, precision of 72.705%, recall of 72.582%, and an F1-score of 72.540%, reflecting a consistent but less effective classification capability. The standout performer is the SBERT-OQA DNN LDA model, which boasts an impressive accuracy, precision, recall, and F1-score all at 94.421%, showcasing exceptional and well-balanced performance across all metrics for the "relevant" column.

Table 1: Overall Performance Comparison of Classification models for column “relevant”.

Algorithm	Accuracy	Precision	Recall	F1-Score
SBERT-OQA SGD Classifier [relevant]	78.133	78.165	78.135	78.127
SBERT-OQA CNB Classifier [relevant]	72.576	72.705	72.582	72.540
SBERT-OQA DNN LDA [relevant]	94.421	94.421	94.421	94.421

The performance comparison of classification models for the "type" column, as presented in Table 2, highlights varying levels of effectiveness. The SBERT-OQA SGD Classifier achieves an accuracy of 94.601%, with precision at 94.493%, recall at 94.554%, and an F1-score of 94.493%, demonstrating strong and consistent performance. The SBERT-OQA CNB Classifier, however, shows a lower accuracy of 81.582%, with precision at 81.227%, recall at 81.435%, and an F1-score of 80.513%, indicating a moderate but less competitive classification ability. The SBERT-OQA DNN LDA model excels with an accuracy of 98.809%, precision of 98.869%, recall of 98.808%, and an F1-score of 98.819%, reflecting outstanding and well-balanced performance across all metrics for the "type" column.

Table 2: Overall Performance Comparison of Classification models for column “type”.

Algorithm	Accuracy	Precision	Recall	F1-Score
SBERT-OQA SGD Classifier [type]	94.601	94.493	94.554	94.493
SBERT-OQA CNB Classifier [type]	81.582	81.227	81.435	80.513
SBERT-OQA DNN LDA [type]	98.809	98.869	98.808	98.819

5. Conclusion

The research developed a comprehensive pipeline for detecting and classifying cyberattack-related tweets, leveraging advanced NLP and machine learning techniques. By utilizing SBERT embeddings for feature extraction, SMOTE for addressing class imbalance, and evaluating multiple classifiers, the SBERT-OQA DNN and LDA model emerged as the standout performer, achieving an impressive accuracy of 94.42% for the binary "relevant" classification and 98.81% for the multi-class "type" categorization. These results signify substantial performance improvements over baseline models, with the SGD Classifier recording 78.13% and 94.60%, and the CNB Classifier at 72.58% and 81.58%, respectively. The integration of DNN and LDA enhanced precision, recall, and F1-scores by effectively capturing semantic nuances in tweet data, leading to more accurate cyber threat identification.

Challenges such as handling noisy tweet data and computational resource constraints were addressed through preprocessing techniques and efficient batch processing in SBERT feature extraction. The pipeline's scalability was improved with cached models, while custom visualizations aided in performance analysis. Future work can focus on integrating advanced transformer-based models and larger cybersecurity datasets to further improve classification accuracy and adaptability. The system can also be extended with real-time threat monitoring capabilities and deployment in large-scale cybersecurity intelligence platforms.

References

[1] Moreno, M.A. Cyberbullying. JAMA Pediatrics 2014, 168, 500.

[2] Bu, S.J.; Cho, S.B. A hybrid deep learning system of CNN and LRCN to detect cyberbullying from SNS comments. In Proceedings of the International Conference on Hybrid Artificial Intelligence Systems, Oviedo, Spain, 20–22 June 2018; Springer: Cham, Switzerland, 2018; pp. 561–572.

[3] Mishra, P.; del Tredici, M.; Yannakoudakis, H.; Shutova, E. Author Profiling for Abuse Detection. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; pp. 1088–1098.

[4] Pavlopoulos, J.; Malakasiotis, P.; Bakagianni, J.; Androutsopoulos, I. Improved Abusive Comment Moderation with User Embeddings. In Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism, Copenhagen, Denmark, 2 May 2017.

[5] Davidson, T.; Warmsley, D.; Macy, M.; Weber, I. Automated hate speech detection and the problem of offensive language. In Proceedings of the International AAAI Conference on Web and Social Media, Montreal, QC, Canada, 15–18 May 2017.

[6] Horvat, M.; Gledec, G.; Leontić, F. Hybrid Natural Language Processing Model for Sentiment Analysis during Natural Crisis. Electronics 2024, 13, 1991. https://doi.org/10.3390/electronics13101991

[7] Silvestri, S.; Islam, S.; Papastergiou, S.; Tzagkarakis, C.; Ciampi, M. A Machine Learning Approach for the NLP-Based Analysis of Cyber Threats and Vulnerabilities of the Healthcare Ecosystem. Sensors 2023, 23, 651. https://doi.org/10.3390/s23020651

[8] Saias, J. Advances in NLP Techniques for Detection of Message-Based Threats in Digital Platforms: A Systematic Review. Electronics 2025, 14, 2551. https://doi.org/10.3390/electronics14132551

[9] Merayo, N.; Vegas, J.; Llamas, C.; Fernández, P. Social Network Sentiment Analysis Using Hybrid Deep Learning Models. Appl. Sci. 2023, 13, 11608. https://doi.org/10.3390/app132011608

[10] Mahmud, T.; Prince, M.A.H.; Ali, M.H.; Hossain, M.S.; Andersson, K. Enhancing Cybersecurity: Hybrid Deep Learning Approaches to Smishing Attack Detection. Systems 2024, 12, 490. https://doi.org/10.3390/systems12110490

[11] Topcu, A.E.; Alzoubi, Y.I.; Elbasi, E.; Camalan, E. SocialMedia Zero-Day Attack Detection Using TensorFlow. Electronics 2023, 12, 3554. https://doi.org/10.3390/electronics12173554

[12] Hamed, S.K.; Ab Aziz, M.J.; Yaakub, M.R. Fake News Detection Model on SocialMedia by Leveraging Sentiment Analysis of News Content and Emotion Analysis of Users' Comments. Sensors 2023, 23, 1748. https://doi.org/10.3390/s23041748

[13] Arora, A.; Arora, A.; McIntyre, J. Developing Chatbots for Cyber Security: Assessing Threats through Sentiment Analysis on Social Media. Sustainability 2023, 15, 13178. https://doi.org/10.3390/su151713178

[14] Raj, C.; Agarwal, A.; Bharathy, G.; Narayan, B.; Prasad, M. Cyberbullying Detection: Hybrid Models Based on Machine Learning and Natural Language Processing Techniques. Electronics 2021, 10, 2810. https://doi.org/10.3390/electronics10222810

[15] Saeed, S.; Suayyid, S.A.; Al-Ghamdi, M.S.; Al-Muhaisen, H.; Almuhaideb, A.M. A Systematic Literature Review on Cyber Threat Intelligence for Organizational Cybersecurity Resilience. Sensors 2023, 23, 7273. https://doi.org/10.3390/s23167273

[16] V Murali Mohan, S Vineetha, G Divya, Rekha Gangula, "Evaluation of Twitter Sentiment towards COVID-19," 2022 Second International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE), Bangalore, India, 2022, pp. 1-5, doi: 10.1109/ICATIECE56365.2022.10047332

[17] RekhaGangula,ChinnakkaSudha, ShashiRekha, Muddham Nirmala,"A Conceptual framework for understanding the role of machine learning in artificial intelligence", International Journal Of Advanced Science And Technology Vol. 29, No. 4s, (2020), Pp. 820-825.

[18] Rekha Gangula, Gayatri Nandam, Chinnakka Sudha, Shashi Rekha ,"Usage of Machine Learning algorithms in DataMining", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878,Volume-8, Issue- 1C2, May 2019

[19] Lingala Thirupathi, Rekha gangula, Sandeep Ravikanti, Jujuroo Sowmya, SK Shruthi"False news Recognition using Machine Learning "Journal of Physics: Conference Series, Volume 2089, 1st International Conference on Applied Mathematics, Modeling and Simulation in Engineering (AMSE) 2021 15-16 September 2021, India (Virtual).

[20] Lingala Thirupathi, G Rekha, SK Shruthi, B Sowjanya, Sowmya Jujuroo," A Novel Twitter Sentimental Analysis Approach Using Naive Bayes Classification", Intelligent System Design, 2023, Volume 494,ISBN : 978-981-19-4862-6.