American Journal of
AI Digital Transformation and Regenerative Pharmacist
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0418 www.zesterapublications.com Original Research Paper

Real-Time Threat Detection and Alerting in Social Streams Using a Dual-Target Transformer Architecture

Gurle Renuka, K. Rakesh*
Vaagdevi Engineering College, Warangal, 506005, Telangana, India.
*Correspondence: K. Rakesh

ABSTRACT

The rapid growth of digital technologies and online communication platforms has significantly increased cybersecurity-related discussions across social media, online forums, and cyber incident reporting platforms. These platforms generate large volumes of unstructured textual data containing valuable information about cyber threats, vulnerabilities, and attack activities. Traditional cybersecurity monitoring methods mainly rely on manual analysis and rule-based filtering techniques, where predefined keywords and pattern-matching approaches are used to identify potential threats. Although these methods are effective for basic monitoring, they become inefficient when handling large-scale datasets and often fail to capture the contextual meaning of textual information. One of the major challenges in cybersecurity intelligence is extracting meaningful insights from complex, noisy, and unstructured text. Conventional approaches struggle with inconsistent language, irrelevant content, and high data volume, reducing the accuracy of cyberattack classification. To address these challenges, this research proposes a machine learning-based framework for automated cyberattack detection using textual data. The framework begins with natural language preprocessing to clean and normalize the text, followed by semantic feature extraction using Sentence-Bidirectional Encoder Representations from Transformers (SBERT) embeddings trained with an Open Question Answering objective. To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) is applied to generate balanced training data. The framework evaluates multiple classification models, including Stochastic Gradient Descent (SGD), Complement Naïve Bayes (CNB), and a hybrid model combining Dense Neural Networks (DNN) with Linear Discriminant Analysis (LDA). Experimental results demonstrate that the hybrid approach improves classification performance, enhancing the accuracy and efficiency of automated cybersecurity text analysis for large-scale cyberattack detection.

Keywords: Cybersecurity, Text Mining, Unstructured Data Analysis, Natural Language Processing, Cyber Threat Intelligence, Feature Extraction, Data Preprocessing, Class Imbalance.

1. INTRODUCTION

The widespread adoption of social media and online communication platforms has transformed the way people exchange information, enabling users to share opinions, news, and real-time updates instantly across the globe [1]. Microblogging platforms and other social networking services have become essential sources of communication due to their speed and accessibility. However, alongside these benefits, they have also emerged as channels for the dissemination of harmful content, including cyber threats, misinformation, malicious campaigns, and coordinated attacks. The openness and accessibility of these platforms allow malicious users to exploit them for spreading fear, promoting cybercriminal activities, and targeting individuals, organizations, and critical infrastructures.

The exponential increase in user-generated content has made manual identification of malicious or suspicious messages impractical [2]. Every minute, millions of posts are published, making it extremely difficult for cybersecurity analysts and platform administrators to monitor online discussions

American Journal of
AI Digital Transformation and Regenerative Pharmacist
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0418 www.zesterapublications.com Original Research Paper

effectively. Consequently, automated systems capable of processing and analyzing textual data in real time have become essential for identifying potential cyber threats. As illustrated in **Fig. 1**, intelligent cyberattack detection frameworks utilize textual information from social media platforms to support timely threat identification and response.

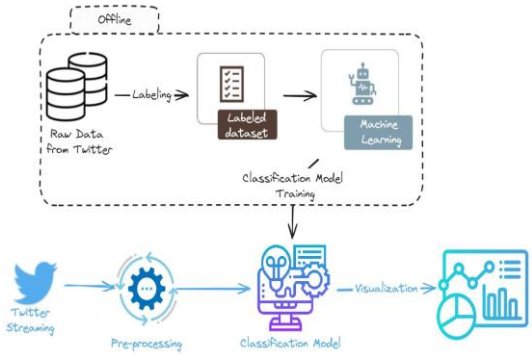


Fig. 1: Overview for twitter-based cyberattack detection.

Natural Language Processing (NLP) has emerged as a powerful technology for analyzing social media text and identifying linguistic patterns associated with malicious or threatening activities [3]. By extracting semantic information and contextual relationships from unstructured textual data, NLP techniques enable automated systems to recognize suspicious content more accurately than conventional keyword-based methods. These approaches improve the ability of cybersecurity systems to process large volumes of online discussions while reducing the dependence on manual monitoring.

Despite these advancements, detecting cyber threats from social media remains a challenging task because of the dynamic and informal nature of online communication [4]. Social media users frequently employ abbreviations, slang, emojis, hashtags, and coded expressions that make it difficult for traditional text analysis techniques to interpret the intended meaning accurately. Moreover, communication patterns evolve rapidly, requiring intelligent detection systems that can continuously adapt to new language styles while maintaining high classification accuracy and computational efficiency.

To overcome these challenges, recent research has focused on integrating advanced Natural Language Processing techniques with machine learning-based cybersecurity frameworks [5]. These hybrid approaches combine semantic text representation, contextual feature extraction, and intelligent classification methods to improve the detection of cyber threats in real time. Such frameworks enhance the capability of security organizations and digital platforms to monitor online environments, identify emerging cyber risks more effectively, and support proactive cybersecurity decision-making.

2. LITERATURE SURVEY

Raj, et al. [6] presented a comprehensive framework for cyberbullying detection by comparing multiple traditional machine learning algorithms with shallow neural network models. Their study also examined the impact of various feature extraction techniques, including TF-IDF and GloVe embeddings, on classification performance. Experimental results showed that Logistic Regression performed best among conventional classifiers, whereas Bi-GRU and Bi-LSTM achieved superior results among neural network models, with an accuracy of approximately 95% and an F1-score close to 98%. Topcu, et al. [7] proposed a Twitter-based cyber-threat monitoring system that utilized TensorFlow for large-scale tweet processing and the Natural Language Toolkit (NLTK) for multilingual keyword extraction. Their approach successfully detected nearly 80% of zero-day attacks, enabling timely vulnerability identification. Atawneh, et al. [8] investigated several deep learning architectures, including CNN, RNN, LSTM, and BERT, for phishing email detection. By applying Natural Language Processing techniques for feature extraction, the proposed framework achieved a maximum detection accuracy of 99.61% using the

American Journal of
AI Digital Transformation and Regenerative Pharmacist
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0418 www.zesterapublications.com Original Research Paper

combined BERT-LSTM model. Arora, et al. [9] introduced an AI-driven chatbot capable of monitoring cybersecurity discussions on social media through sentiment analysis. The chatbot continuously analyzed Twitter conversations to predict potential cyber threats and provided valuable insights for improving organizational cybersecurity awareness and proactive threat management.

Merayo, et al. [10] proposed a hybrid deep learning model combining Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks for sentiment analysis of Spanish Twitter messages. The model effectively captured both local textual features and contextual information, achieving approximately 76% accuracy for three-class sentiment classification and 91% accuracy for binary classification. Koloveas, et al. [11] developed an integrated cyber-threat intelligence platform named inTIME that supports the complete cyber-threat intelligence lifecycle. The framework automatically gathers information from social media platforms, online forums, security repositories, marketplaces, and hidden web resources using intelligent crawlers and monitoring services. It further incorporates natural language understanding, entity correlation, and collaborative intelligence management to support cybersecurity analysts in collecting, processing, and sharing actionable threat intelligence. Hamed, et al. [12] proposed a Bidirectional Long Short-Term Memory (Bi-LSTM) framework for fake news detection by combining textual content with sentiment analysis of news articles and emotion analysis of user comments. The proposed approach achieved an Area Under the ROC Curve (AUC) score of 96.77%, demonstrating that integrating multiple contextual features significantly enhances detection performance.

Saias, et al. [13] conducted a systematic review of Natural Language Processing techniques for detecting message-based threats in digital communication by analyzing thirty research publications following the PRISMA methodology. Their study examined commonly used NLP techniques, evaluation methods, and ethical considerations, concluding that large language models remain underutilized while privacy-related challenges require greater attention in future research. Mahmud, et al. [14] introduced a hybrid deep learning architecture that integrates Convolutional Neural Networks (CNN) with Bidirectional Gated Recurrent Units (Bi-GRU) for SMS phishing detection. The framework employed Word2Vec embeddings to transform textual messages into semantic feature vectors before classification. Experimental evaluation on the SMS Phishing Collection dataset demonstrated an overall detection accuracy of 99.82%, outperforming several existing machine learning and deep learning approaches. Coyac-Torres, et al. [15] presented an approach based on natural language processing tools and a convolutional neural network architecture to detect and classify four types of cyberattacks in social network messages, including malware, phishing, spam, and even one whose aim is to deceive a user into spreading malicious messages to other users, which, in this work, is identified as a bot attack.

3. PROPOSED SYSTEM

The proposed system architecture is designed to automatically detect and classify cyberattack-related tweets using an intelligent Natural Language Processing (NLP) and machine learning pipeline. Initially, textual data is collected from multiple sources and transformed into a structured format suitable for analysis. The collected data then undergoes preprocessing to remove unwanted content and improve text quality before semantic feature extraction is performed using Sentence-BERT (SBERT) embeddings. The extracted features are supplied to multiple classification models, including SGD, CNB, and a Dense Neural Network (DNN), to learn meaningful attack patterns. The deep features generated by the DNN are further processed using Linear

American Journal of
AI Digital Transformation and Regenerative Pharmacist
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0418 www.zesterapublications.com Original Research Paper

Discriminant Analysis (LDA) to improve class separability and enhance prediction performance. As illustrated in Fig. 2, the outputs from different models are integrated through a Flask-based decision module, which produces the final cyberattack classification results. This architecture improves classification accuracy, efficiently handles large-scale textual data, and supports scalable real-time cybersecurity threat detection.

Step 1: Data Ingestion

The system begins by collecting cyberattack-related textual data from Twitter and Watson NLU datasets. The raw tweets are parsed and converted into a structured format so that they can be processed consistently. This stage establishes a reliable input source for all subsequent machine learning operations.

Step 2: Text Preprocessing

The collected text is cleaned using Natural Language Processing techniques such as tokenization, lemmatization, stemming, stop-word removal, and text normalization. These operations eliminate irrelevant information and reduce textual noise, producing standardized and meaningful input data for feature extraction.

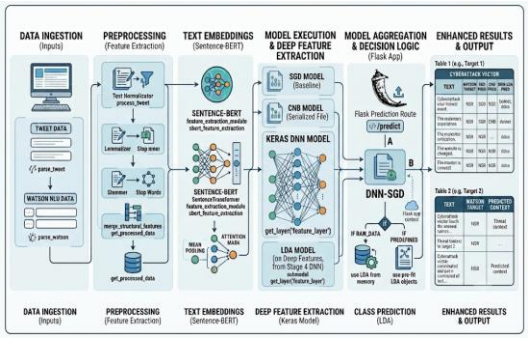


Fig. 2: System architecture of threat detection and alerting in social streams

Step 3: Semantic Feature Extraction Using SBERT

The preprocessed text is transformed into dense semantic vector representations using Sentence-BERT (SBERT). Mean pooling and attention masking are applied to generate contextual sentence embeddings that preserve the semantic meaning of each tweet. These embeddings provide high-quality features for classification.

Step 4: Model Training and Deep Feature Extraction

The extracted SBERT embeddings are evaluated using multiple classifiers, including Stochastic Gradient Descent (SGD), Complement Naïve Bayes (CNB), and a Dense Neural Network (DNN). The DNN learns complex hidden representations from the embedding vectors, and the deep feature layer is extracted for further analysis using Linear Discriminant Analysis (LDA).

Step 5: Decision Aggregation and Classification

The deep features produced by the DNN are processed using LDA to improve class discrimination and prediction accuracy. A Flask-based decision module integrates the outputs from different classifiers and generates the final cyberattack category. The classified results are presented in tabular form, providing accurate and efficient cyberattack detection for large-scale social media data.

4. RESULTS ANALYSIS

The results and description phase focuses on evaluating the performance of the cyberattack detection framework after the successful implementation of preprocessing, feature extraction, and machine learning models. In this stage, the trained models are tested using unseen data to analyze their ability to correctly classify cybersecurity-related textual information. The prediction results generated by different machine learning models are compared to understand their effectiveness in detecting cyberattack categories and relevance. Performance evaluation metrics such as accuracy, precision, recall, and F1-score are used to measure the efficiency of each model. Visualization techniques are also applied to present the comparison of model performance through graphs and charts. These results provide insights into how well the analytical framework processes textual data and identifies cyberattack patterns. The analysis of

American Journal of
AI Digital Transformation and Regenerative Pharmacist
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0418 www.zesterapublications.com Original Research Paper

experimental results helps determine the most effective model for cybersecurity text classification and supports the validation of the proposed hybrid learning approach.

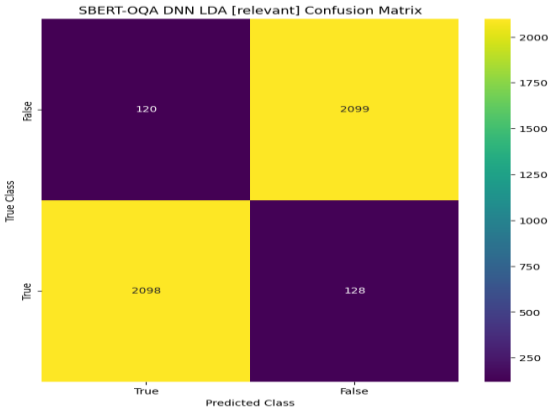


Fig. 3: Confusion matrix obtained SBERT-OQA DNN+LDA. for column relevant.

The SBERT-OQA DNN+LDA matrix as shown in the Fig. 3, 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.

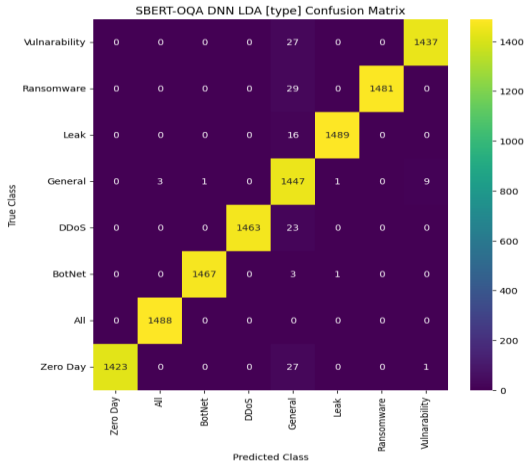


Fig. 4: Confusion matrix obtained using SBERT-OQA DNN+LDA. for column “type”.

The SBERT-OQA DNN+LDA matrix as shown in the Fig. 4 for "type" shows strong performance, with 1,437 correct predictions for "Vulnerability," 1,481 for "Ransomware," 1,469 for "Leak," and 1,423 for "Zero Day." The low off-diagonal values (e.g., 27 for "Vulnerability" as "Ransomware") indicate high accuracy and minimal confusion, leveraging the deep learning and LDA approach effectively across the 4,445 instances.

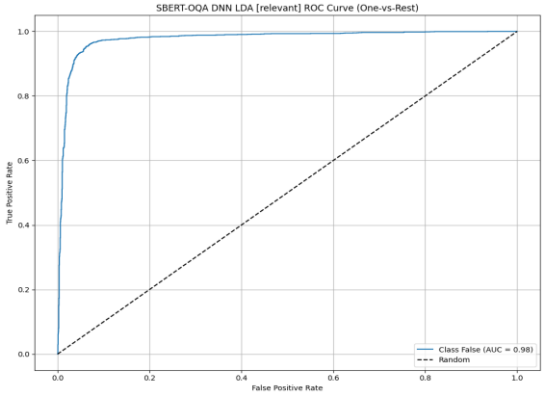


Fig. 5: ROC curve obtained using SBERT-OQA DNN+LDA. for column “relevant”.

The SBERT-OQA DNN+LDA ROC curve demonstrates as shown in the Fig. 5, the highest performance, with an AUC of 0.98. The curve closely follows the top-left corner, indicating excellent sensitivity and specificity for the "relevant" column, outperforming the other models.

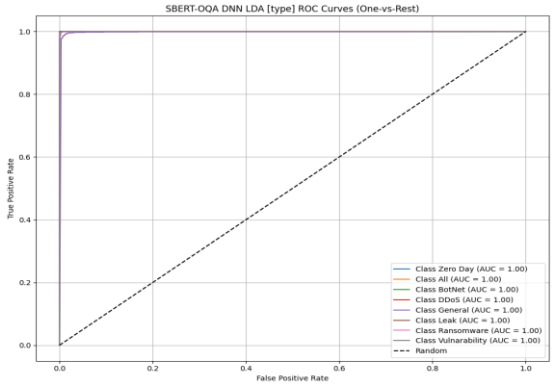


Fig. 6: ROC curve obtained using SBERT-OQA DNN+LDA. for column “type”.

The SBERT-OQA DNN+LDA ROC curves for "type" exhibit perfect AUCs of 1.00 for all classes ("Zero Day," "All," "BotNet," etc.) as shown in the Fig. 6. The curves align closely with the top-left corner, demonstrating exceptional discriminative power and near-perfect classification accuracy across all categories for the "type" column.

Comparative Analysis

American Journal of
AI Digital Transformation and Regenerative Pharmacist
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0418 www.zesterapublications.com Original Research Paper

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

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

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.

5. CONCLUSION

The proposed research developed a comprehensive machine learning framework for detecting and classifying cyberattack-related tweets by integrating advanced Natural Language Processing (NLP) and machine learning techniques. The framework employed Sentence-Bidirectional Encoder Representations from Transformers (SBERT) embeddings for semantic feature extraction and the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance within the dataset. Multiple classification models were evaluated to determine the most effective approach. Among the evaluated models, the hybrid SBERT-OQA DNN-LDA model achieved the best performance, attaining an accuracy of 94.42% for binary relevance classification and 98.81% for multi-class cyberattack type classification. In comparison, the Stochastic Gradient Descent (SGD) classifier achieved accuracies of 78.13% and 94.60%, while the Complement Naïve Bayes (CNB) classifier obtained 72.58% and 81.58% for the respective classification tasks. The integration of Dense Neural Networks (DNN) with Linear Discriminant Analysis

American Journal of
AI Digital Transformation and Regenerative Pharmacist
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0418 www.zesterapublications.com Original Research Paper

(LDA) enabled the framework to capture complex semantic relationships within tweet data, resulting in improved precision, recall, and F1-score compared with conventional machine learning models. Furthermore, challenges associated with noisy social media text and class imbalance were effectively addressed through comprehensive preprocessing techniques, efficient SBERT-based feature extraction, and balanced data generation using SMOTE. The use of cached embedding models and optimized batch processing also enhanced computational efficiency and scalability. Overall, the proposed framework provides an accurate, scalable, and effective solution for automated cyberattack detection and classification from large-scale social media text.

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] 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

[7] 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

[8] Atawneh, S.; Aljehani, H. Phishing Email Detection Model Using Deep Learning. Electronics 2023, 12, 4261. https://doi.org/10.3390/electronics12204261

[9] 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

[10] 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

[11]Koloveas, P.; Chantzios, T.; Alevizopoulou, S.; Skiadopoulos, S.; Tryfonopoulos, C. inTIME: A Machine Learning-Based Framework for Gathering and Leveraging Web Data to Cyber-Threat Intelligence. Electronics 2021, 10, 818. https://doi.org/10.3390/electronics10070818

[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]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

[14]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

[15]Coyac-Torres, J.E.; Sidorov, G.; Aguirre-Anaya, E.; Hernández-Oregón, G. Cyberattack Detection in Social Network Messages Based on Convolutional Neural Networks and NLP Techniques. Mach. Learn. Knowl. Extr. 2023, 5, 1132-1148. https://doi.org/10.3390/make5030058x