Dual-Target Learning with Transformer-Based OQA for Real-Time Social Threat Analysis

K. Chiranjeevi[1], Y. Mohan Krishna[2], U. Bharath[2], Vansh Naidu[2], V. Jayanth[2]

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

[1,2]Geethanjali Institute of Science and Technology, Nellore-Bombay Highway, S.P.S.R, Andhra Pradesh 524137, India

ABSTRACT

The rapid growth of social networking platforms such as Twitter has significantly increased the exposure of users and servers to complex cyber-attacks, including botnet-driven campaigns, phishing, distributed denial-of-service (DDoS) attacks, vulnerability exploitation, and information leakage. Traditional OSI-layer, rule-based security mechanisms suffer from high time complexity, limited adaptability, and an inability to detect multiple coordinated attacks in real time. To address these limitations, this research proposes a Dual-Target Transformer-Driven Optimized Question Answers (OQA) framework for real-time cyber threat detection and prediction in Twitter streams. The proposed system is designed with two outputs: Primary Output performs accurate cyber-attack type classification, while Secondary Output determines the relevance and severity of the detected threat for intelligent prioritization and response. The framework utilizes a large-scale Twitter cyber-security dataset and integrates advanced NLP preprocessing, including contextual normalization, noise removal, semantic enrichment, along with Uniform Resource Locator (URL)-specific feature extraction and destination-based preprocessing to analyse domain reputation, redirection behaviour, and landing page characteristics. A Transformer-based learning architecture using Sentence Bidirectional Encoder Representations from Transformers (SBERT) combined with the OQA algorithm is employed for deep semantic feature learning, supported by Synthetic Minority Over-sampling Technique (SMOTE)-based data balancing and Stochastic Gradient Descent (SGD) optimization for efficient training under imbalanced data conditions. To further enhance detection performance, the system is complemented with Naïve Bayes, Deep Neural Networks (DNN), and Linear Discriminant Analysis (LDA) for probabilistic validation, deep behavioural learning, and topic-level attack interpretation. Experimental results demonstrate that the proposed dual-target hybrid framework achieves high accuracy, low false alarm rates, and robust real-time detection capability for multiple cyber-attack types on Twitter. The outcomes validate the effectiveness of the proposed system as a scalable, intelligent, and adaptive cyber defence solution for modern social network environments.

Key words: Cyber Threat Detection, Transformer Models, SBERT, Optimized Question Answering (OQA), Twitter Streams, Attack Classification, Threat Severity Analysis, Real-Time Security.

1. INTRODUCTION

Cyber-attacks on Twitter manifest in multiple forms. Phishing campaigns are among the most prevalent, often using shortened malicious links and impersonation tactics to lure users into divulging credentials. Attackers frequently compromise verified or high-visibility accounts to increase the credibility and reach of their deceptive messages. Malware delivery through deceptive advertisements and malicious Uniform Resource Locators (URLs) is another significant threat, leading to widespread system compromise and data theft. In addition, bot-driven coordinated harassment campaigns target individuals and organizations, resulting in psychological harm and reputational damage. Another major threat in the Twitter ecosystem is bad propaganda and disinformation. Unlike traditional cyber-attacks that primarily target systems and data, propaganda attacks target human cognition and social trust.

Figure 1 illustrates one of the most significant security incidents in Twitter's history, where multiple high-profile accounts were compromised on July 15, 2020. The chart lists several influential public figures such as including Barack Obama, Kim Kardashian, Bill Gates, Elon Musk, and others along with their follower counts at the time of the attack. These individuals represent some of the most widely followed accounts on the platform, which provided the attackers with immediate access to a massive global audience. Once inside these accounts, the attackers posted fraudulent messages promoting a cryptocurrency scam, which urged users to send Bitcoin to a specific wallet under false promises. The wide reach of these profiles increased the probability of rapid spread and high user engagement with the malicious content.

The figure demonstrates how a single coordinated breach can create extensive impact across the platform, especially when verified accounts are involved. The compromise of influential figures reflects weaknesses in account protection mechanisms and highlights the consequences of insufficient monitoring of suspicious access patterns. This incident triggered significant concern regarding Twitter's internal security controls, user authentication processes, and vulnerability management. The event also revealed how attackers can exploit trusted online identities to mislead large populations, influence user behaviour, and carry out financial theft. Overall, the figure emphasizes the critical need for stronger cybersecurity practices and continuous monitoring to prevent unauthorized access and protect high-value digital assets.

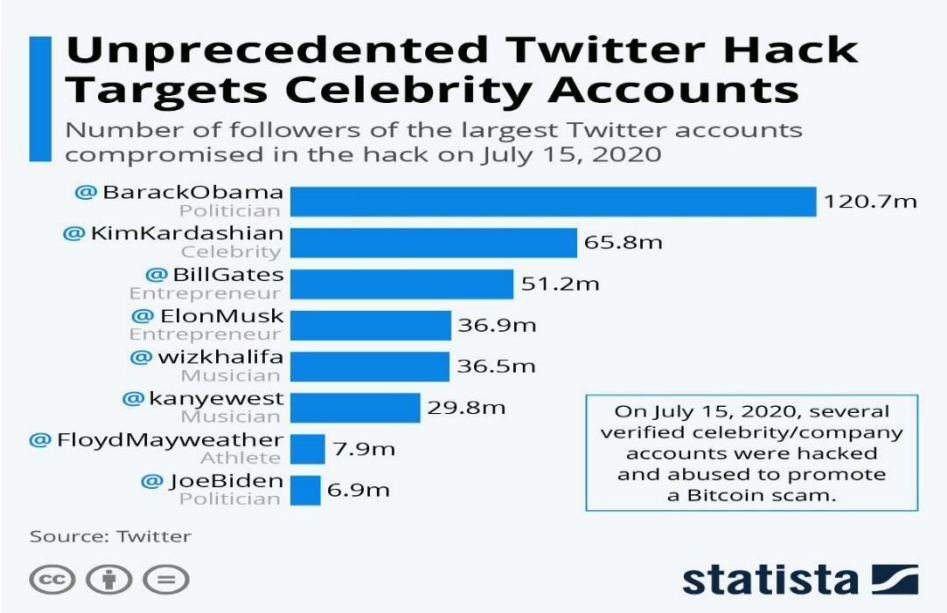

Figure 1: Unprecedented Twitter Hack Targets Celebrity Accounts

Figure 2 presents a comparison of how social media users and Twitter users perceive the security of their personal information after the major Twitter hack. The chart shows consistently high levels of distrust across key data categories such as geolocation, full name, email address, birth date, and banking information. Geolocation appears as the most sensitive category, with 75% of social media users and 69% of Twitter users expressing little or no confidence in platform security. Similar patterns appear across all data types, indicating widespread concern about weak protection measures and the growing exposure of personal data to cyber threats. These results reflect strong user awareness of the risks associated with online platforms and the consequences of security failures. The figure also demonstrates that this breach triggered broader concerns beyond the Twitter community, influencing how users evaluate the overall reliability of social media platforms. Although distrust levels are slightly higher among general social media users, a significant portion of Twitter users also report reduced confidence

in the platform's ability to safeguard sensitive information. This decline in trust underscores the urgent need for stronger security mechanisms, improved access control, and more transparent data protection practices. The trends captured in this figure highlight the importance of building resilient systems that can handle increasing threats, fortify user privacy, and restore public confidence in digital communication environments.

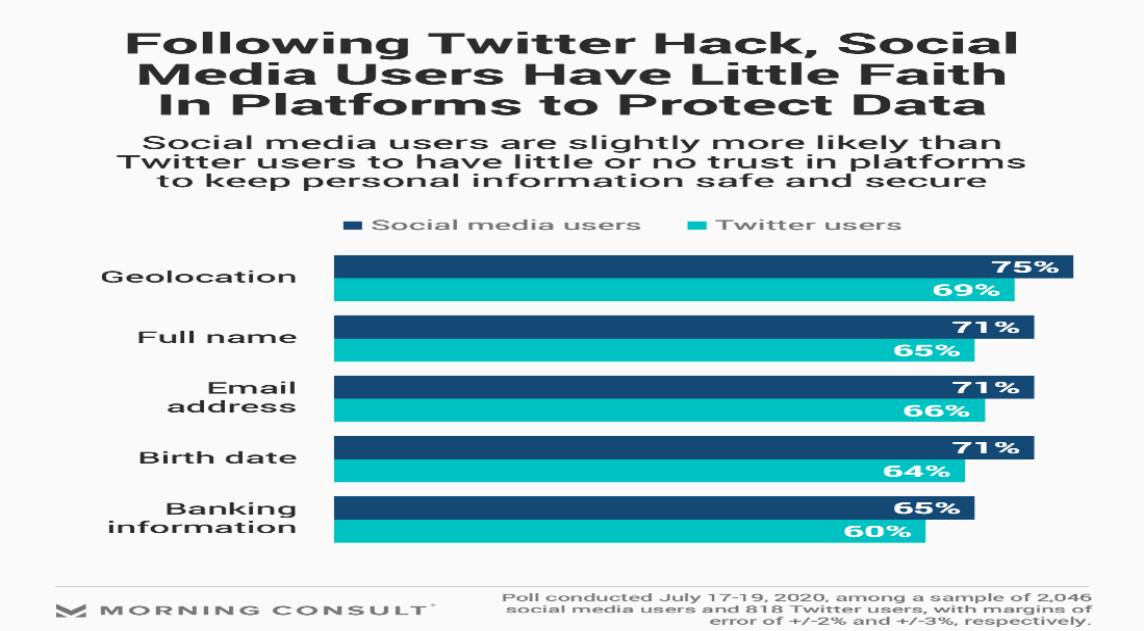

Figure 2: Faith issues on Social Media Platforms

2. LITERATURE SURVEY

Kundiya, et al. [1] systematic review, conducted under the PRISMA 2020 framework, investigates the application of Natural Language Processing (NLP) in insider threat detection, integrating Machine Learning (ML) and Deep Learning (DL) techniques, based on literature from 2019 to 2024. Addressing research questions (RQ1–RQ5) on techniques, datasets, performance metrics, challenges, and future directions, and evaluated via quality assessment criteria (QA1–QA4), the study synthesizes 66 high-quality studies from an initial 132 records across databases like IEEE, ScienceDirect, and Scopus. Findings reveal a surge in publications, peaking in 2023 and 2024, with Deep Learning (30.3%), hybrid NLP-ML/DL (27.3%), traditional ML (27.3%), and NLP-only (15.2%) approaches dominating, leveraging datasets like CERT and Enron to achieve accuracies up to 99.2% and F1-scores above 94%. However, reliance on synthetic data limits real-world applicability. Challenges include model complexity, explainability issues, binary classification biases, and dataset gaps, while future opportunities lie in lightweight, interpretable models, hybrid pipelines, anonymized datasets, and digital twins.

Priyanka, et al. [2] proposed an Intelligent NLP-Driven Framework designed to automatically detect and profile cyber threats by extracting, analysing, and interpreting threat-related information from unstructured textual data sources such as social media, dark web forums, and cybersecurity reports. Leveraging advanced Natural Language Processing (NLP) and machine learning models, the framework identifies key threat indicators, attack patterns, and entity relationships in real time. The system aims to enhance the efficiency of cyber threat intelligence (CTI) generation while reducing analyst workload. Experimental evaluation demonstrates improved accuracy in early detection and profiling compared with baseline keyword-based systems, establishing the framework as a scalable and adaptive solution for next-generation cyber defence.

Kovalchuk, et al. [3] achieved this goal, machine learning and natural language processing techniques were employed, particularly the adaptation of large language models for threat classification, risk-level evaluation, and anomaly detection. A system was developed to analyse incoming and outgoing email communications, which during testing automatically identified phishing attacks and social engineering techniques, assigned risk scores to messages, and quarantined those exceeding a predefined threshold (e.g., 0.8) for further inspection. The system analysed a dataset of 100,000 emails, of which 70% were legitimate communications and 30% were phishing attacks. Additionally, real-time analysis of data streams from corporate logs and external sources enabled the detection of potential cyber incidents with an accuracy of up to 94%, while reducing the false-positive rate to 6.5%. The obtained results confirmed the efficacy of large language models, which achieved a threat classification accuracy of up to 97% with an F1-score of 95% and reduced incident response times by 30-40%.

Pope, et al. [4] presented Social Cipher, an AI-driven architecture for preventive threat detection that integrates a custom threat framework, robust filtering mechanisms, and Sent Intel GPT (4.0), a transformer model optimized for multimodal analysis. Social Cipher categorizes threats across six dimensions-cyber, physical, reputational, operational, regulatory, and psychological/emotional-enabling risk scoring aligned with government classification levels. Empirical evaluations of Sent Intel GPT against ten contemporary AI models (including GPT 3.5, GPT 4.0, Claude, and Mistral variants) show it achieves 96 % accuracy, confirming its ability to interpret complex inputs. These findings highlight Social Cipher's potential to improve real-time threat detection and support proactive interventions in dynamic digital ecosystems.

Lipianina-Honcharenko, et al. [5] obtained results confirm all five proposed hypotheses: the developed sentiment analysis module achieves macro-F1 = 0.85 and reduces MAE by 18.2% compared to the baseline model; the polarity inversion detection algorithm allows automatic reversal of sentiment score in manipulative texts, improving the detection of hostile narratives; the hybrid thematic classification achieves macro-F1=0.83, with latency of 55 m s /document and throughput of 18 documents/second; integration of all modules into a unified pipeline improves recall by 10.4% without significant increase in latency; the RAIE conceptual model ensures $\Delta F1 \leq 5\%$, an expert user satisfaction score of 4.14/5 and less than 10% latency overhead.

Ibitoye, et al. [6] explored a comprehensive AI-enhanced threat modelling framework that synchronizes multi-domain surveillance data streams for predictive intelligence and real time response. It proposes a unified architecture leveraging machine learning algorithms, computer vision, and natural language processing (NLP) to identify anomalous behaviours, detect intent, and model risk trajectories across cyber intrusions, aerial incursions, maritime violations, and ground-based threats. By fusing sensor data from satellites, drones, radar, SONAR, and digital telemetry, the system creates a holistic threat landscape capable of early-warning and preventive mitigation. The paper further examines domain-specific use cases, such as AI-driven unmanned aerial surveillance in border security, behavioural analytics for cyber threat actor profiling, and deep learning techniques for autonomous maritime anomaly detection. Validation is supported by cross-domain data sets and stress-tested simulation environments to benchmark performance, accuracy, and response latency.

Potla, et al. [7] proposed model combines BERT for text class, ResNet50 for photograph processing, and a hybrid LSTM-3-d CNN community for video content material analysis. We constructed a large-scale dataset comprising 500,000 textual posts, 200,000 offensive images, and 50,000 annotated motion pictures from more than one platform, which includes Twitter, Reddit, YouTube, and online gaming forums. The system became carefully evaluated using trendy gadget mastering metrics which include accuracy, precision, remember, F1-score, and ROC-AUC curves. Experimental outcomes demonstrate

that our multi-modal method extensively outperforms single-modal AI classifiers, achieving an accuracy of 92.3%, precision of 91.2%, do not forget of 90.1%, and an AUC rating of 0.95.

Reza, et al. [8] searchers used the Hornet 40 dataset which includes network traffic collected over the course of 40 days from honeypots in eight places: Amsterdam, London, Frankfurt, San Francisco, New York, Singapore, Toronto, and Bangalore. Various machine learning approaches are used within a data-driven system to spot and detect abnormal traffic and threats in the network such as Random Forest, Support Vector Machines (SVM), Long Short-Term Memory (LSTM) networks and Isolation Forests. At the same time, data, and findings from public threat intelligence, darknet sources and cybersecurity forums are studied using Natural Language Processing (NLP) to find important information about threats. As a result of this, the detection rate is improved by comparing suspicious traffic in honeypots with global findings and the reported IOCs. Combining AI and OSINT together allows the engine to read and analyse a lot of network data quickly and in almost real time.

Vyas, et al. [9] research paper explores the integration of AI technologies into the risk infrastructure of global financial institutions, focusing on predictive analytics, anomaly detection, real-time decision-making, and risk scoring. By examining current applications in institutions like JPMorgan Chase and PayPal, we evaluate how AI has reshaped practices and policy frameworks across the United States, Europe, and emerging economies in Asia and Africa.

Odunaike, et al. [10] study begins with an overview of traditional risk modelling limitations, particularly in the face of black swan events, flash crashes, and sector-specific anomalies. It then presents an architecture for dynamic risk modelling that incorporates real-time data pipelines, data normalization layers, and AI driven analytics engines. Emphasis is placed on the integration of machine learning algorithms capable of adapting to new data patterns, identifying emerging risk clusters, and recalibrating portfolio exposures accordingly. Techniques such as online learning, temporal convolutional networks, and ensemble forecasting models are highlighted for their robustness and adaptability.

Umechukwu, et al. [11] proposed framework follows a four-phase architecture: data collection, data analysis, risk assessment, and mitigation, with a continuous feedback loop for adaptive learning. Experiments were conducted using LU Flow and IEC 60870–5-104 intrusion detection datasets, yielding a detection accuracy of approximately 99.96% and significant improvements in precision, recall, and F1-score compared with baseline models such as Security BERT, XAI, and DNN-based solutions.

Jabed, et al. [12] proposed an integrated framework that leverages Business Process Intelligence (BPI) and advanced Artificial Intelligence (AI) analytics to detect threats in real time, aligned with the unique demands of these sectors. The framework uses process mining, anomaly detection, deep learning models, and continuous monitoring of event logs to spot the behavioral deviations in the live business processes.

Sangher, et al. [13] utilized research work advanced and lighter version of BERT and LSTM model used yielding accuracy of 90.12% and 91.35% respectively. LSTM performed best to extract multiclass classification of actual intension of social media usage by intelligent analysis on hackers' discussions. Strategies on social media platforms such as Facebook, twitter, Instagram, Snapchat to exploit them using darknet platforms also explored.

Alhady, et al. [14] utilized on public datasets such as CIC-MalMem-2022, CICIDS-2017, and a phishing corpus, the proposed model demonstrates high efficacy, achieving an average F1-score of 98.2% for malware classification, 97.5% for phishing detection, and 96.8% in anomaly-based breach detection with a low false positive rate.

Vijayan, et al. [15] proposed system consists of secure data ingestion pipelines, distributed real time processing using Apache Kafka and Spark Streaming, and an ensemble machine learning based anomaly detection model. A cross-industry use case highlights the framework's ability to detect insider threats in HR systems, fraudulent financial transactions, and cyber espionage in government networks. The system is implemented on AWS cloud infrastructure, and experimental results show that it achieves 98.6% precision for HR insider threat detection, 94.2% accuracy in financial fraud prevention, and a sub-2-second detection latency for public sector security alerts.

3. PROPOSED SYSTEM

The proposed system focuses on automated threat detection from social media streams, specifically Twitter data, by combining natural language processing, deep learning, and statistical topic modelling. The system ingests tweet-level data containing textual content, metadata, and security-related attributes, preprocesses and analyses the data to uncover patterns, and extracts semantic features using SBERT-based open question answering representations. As shown as figure 4 To address class imbalance, SMOTE is applied before training multiple models, including existing machine learning classifiers and a proposed deep neural network integrated with LDA for enhanced topic-aware threat classification. The system classifies tweets into cyber threat categories such as botnet, ddos, ransomware, vulnerability, leak, Zero Day, general, and All, and is deployed through a Flask-based web application for real-time threat prediction.

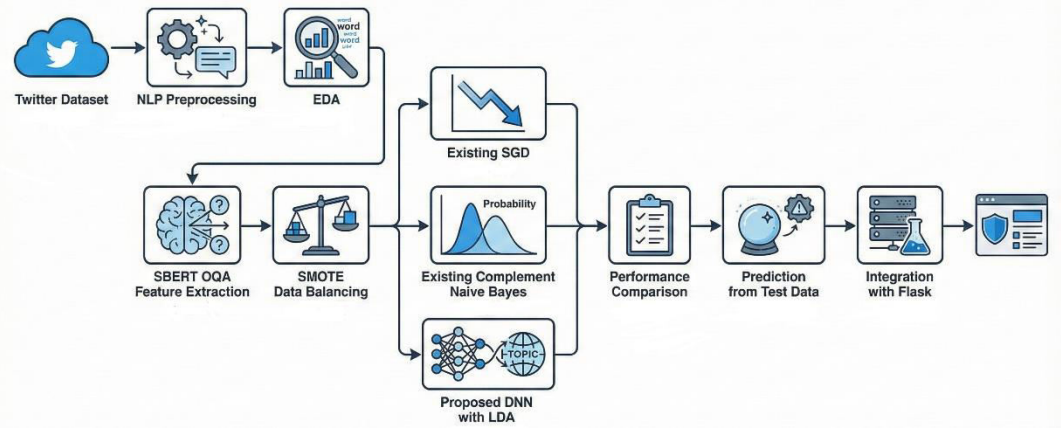

Figure 4: System Architecture

Step 1: Twitter Dataset: In this step, a labelled Twitter dataset is collected containing fields such as relevant text, tweet type, annotations, URLs, destination URLs, Watson analysis outputs, and certificate validity. This dataset serves as the primary input source for threat detection and includes both benign and malicious cyber-related tweets corresponding to different threat categories.

Step 2: NLP Preprocessing: The raw tweet text is cleaned and normalized using NLP preprocessing techniques. This includes removal of URLs, mentions, hashtags, special characters, and stop words, followed by tokenization and lemmatization. The preprocessing step ensures noise reduction and improves the quality of textual input for downstream feature extraction and modelling.

Step 3: EDA: Exploratory Data Analysis is performed to understand the dataset characteristics. This includes analysing class distribution, tweet length statistics, word frequency analysis, and visualization of threat categories. EDA helps identify data imbalance, common threat-related terms, and trends within the social media data.

Step 4: SBERT OQA Feature Extraction: Semantic features are extracted using Sentence-BERT with Open Question Answering representations. SBERT converts each tweet into dense vector embeddings

that capture contextual and semantic relationships between words, enabling more accurate detection of nuanced cyber threat information present in social media text.

Step 5: SMOTE Data Balancing: To address class imbalance in the dataset, Synthetic Minority Over-sampling Technique is applied. SMOTE generates synthetic samples for underrepresented threat classes, ensuring balanced training data and improving model generalization across all cyber threat categories.

Step 6: Existing SGD: An existing Stochastic Gradient Descent classifier is trained using the balanced SBERT feature vectors. This model serves as a baseline machine learning approach for threat detection and provides comparative performance metrics such as accuracy, precision, recall, and F1-score.

Step 7: Existing Complement Naive Bayes: Complement Naive Bayes is implemented as another baseline classifier, particularly suited for imbalanced text classification tasks. It evaluates probabilistic word distributions across threat categories and serves as a comparative benchmark against both SGD and the proposed model.

Step 8: Proposed DNN with LDA: The proposed methodology integrates a Deep Neural Network with Latent Dirichlet Allocation. LDA is used to extract latent threat topics from tweet text, which are combined with SBERT embeddings and fed into the DNN. This hybrid approach enhances threat detection by leveraging both semantic understanding and topic-level contextual awareness.

Step 9: Performance Comparison: The performance of SGD, Complement Naive Bayes, and the proposed DNN with LDA is evaluated using standard metrics such as accuracy, precision, recall, F1-score, and confusion matrices. Comparative analysis demonstrates the effectiveness of the proposed model over existing approaches.

Step 10: Prediction from Test Data: The trained models are applied to unseen test data to predict cyber threat categories. This step validates model robustness and ensures accurate classification of real-world Twitter streams into predefined threat classes.

Step 11: Integration with Flask: The final trained model is integrated into a Flask-based web application. This enables real-time threat detection by accepting tweet input, performing preprocessing and feature extraction, and returning predicted threat categories through a user-friendly interface.

4. RESULTS ANALYSIS

Figure 5 shows the WordCloud representation of the top 100 most frequent words in the social stream dataset used for threat detection. The most dominant term is “http”, indicating the high prevalence of URLs and external links in tweets, which is characteristic of cyber threat reporting and information sharing. Labels such as “False” and “True” appear prominently, reflecting the importance of relevance annotation in the dataset. Security-related terms including “security”, “vulnerability”, “malware”, “ransomware”, “ddos”, “botnet”, and “cybersecurity” are frequently observed, confirming that the dataset is strongly centered around cyber threat discourse. Additional contextual terms such as “govt”, “network”, “attack”, “system”, and “data” further highlight discussions related to infrastructure, governance, and cyber incidents. Overall, the distribution of high-frequency terms demonstrates that the dataset captures both technical threat indicators and contextual information essential for effective threat detection in social streams.

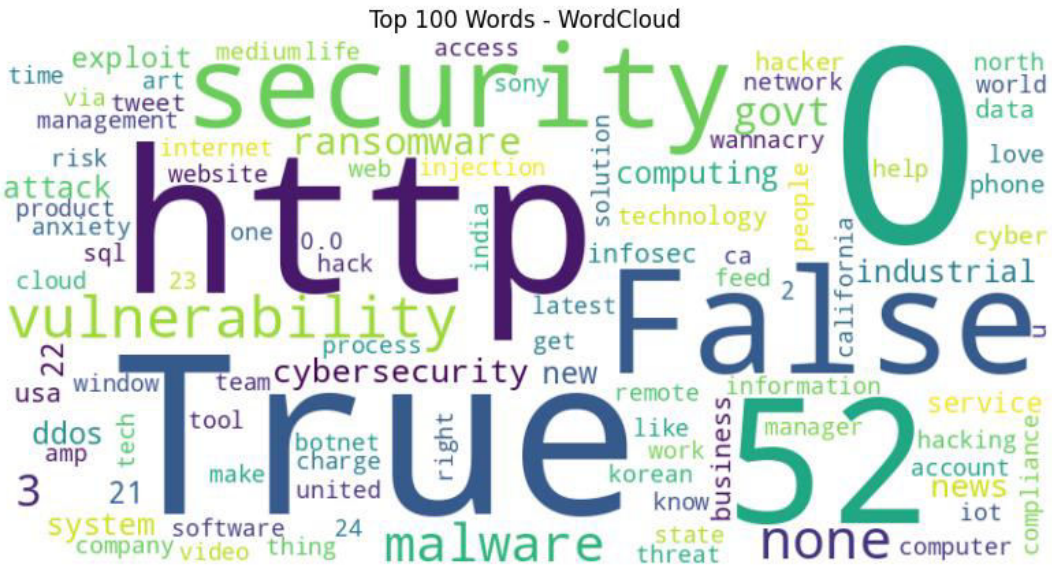

Figure 5: Top 100 Words -Wordcloud

Figure 6 shows the confusion matrix of the SBERT–OQA Complement Naive Bayes classifier for multi-class cyber threat type classification, illustrating both correct predictions and misclassification patterns across attack categories. High diagonal values indicate strong performance for several classes, including Zero Day with 1,407 correct predictions, Vulnerability with 1,434, General with 1,408, Leak with 1,370, and DDoS with 1,263 correctly classified instances. The BotNet class achieves 1,105 correct predictions, but also shows misclassifications into Ransomware (62) and All (165). The Ransomware class exhibits comparatively lower accuracy, with 565 correct predictions, and significant confusion with BotNet (233) and All (237), indicating semantic overlap with other threat categories. The All class records 1,104 correct predictions while being misclassified into Ransomware (67) and Vulnerability (76). Overall, the matrix highlights that while Complement NBC performs well for well-defined categories, it experiences increased confusion among closely related cyber threat types compared to the SGD-based model.

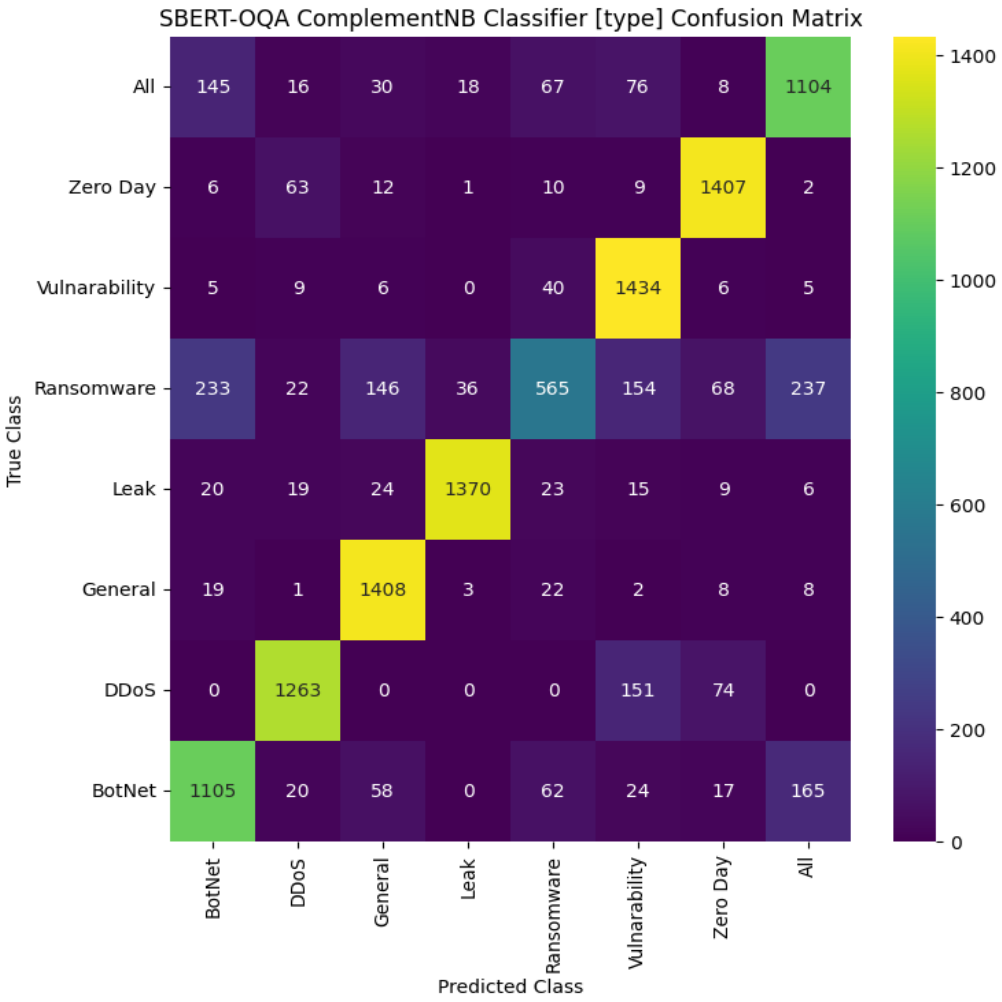

Figure 6: SBERT- OQA Complement NBC Classifier Confusion Matrix

Figure 7 shows the confusion matrix of the SBERT–OQA DNN–LDA classifier for multi-class cyber threat type classification, demonstrating very strong predictive performance across all attack categories. The diagonal values indicate near-perfect classification, with DDoS achieving 1,488 correct predictions, General 1,464, Leak 1,473, Ransomware 1,450, Vulnerability 1,493, Zero Day 1,489, BotNet 1,422, and All 1,449 correctly classified instances. Only minimal misclassifications are observed, such as BotNet being confused with Ransomware (22 cases) and All (7 cases), and Ransomware misclassified as DDoS (3 cases) or General (1 case). The All class shows very limited confusion, with 15 instances misclassified as Ransomware. The figure highlights the effectiveness of combining SBERT semantic embeddings with deep neural feature extraction and LDA, resulting in highly discriminative representations and significantly reduced class overlap compared to classical models.

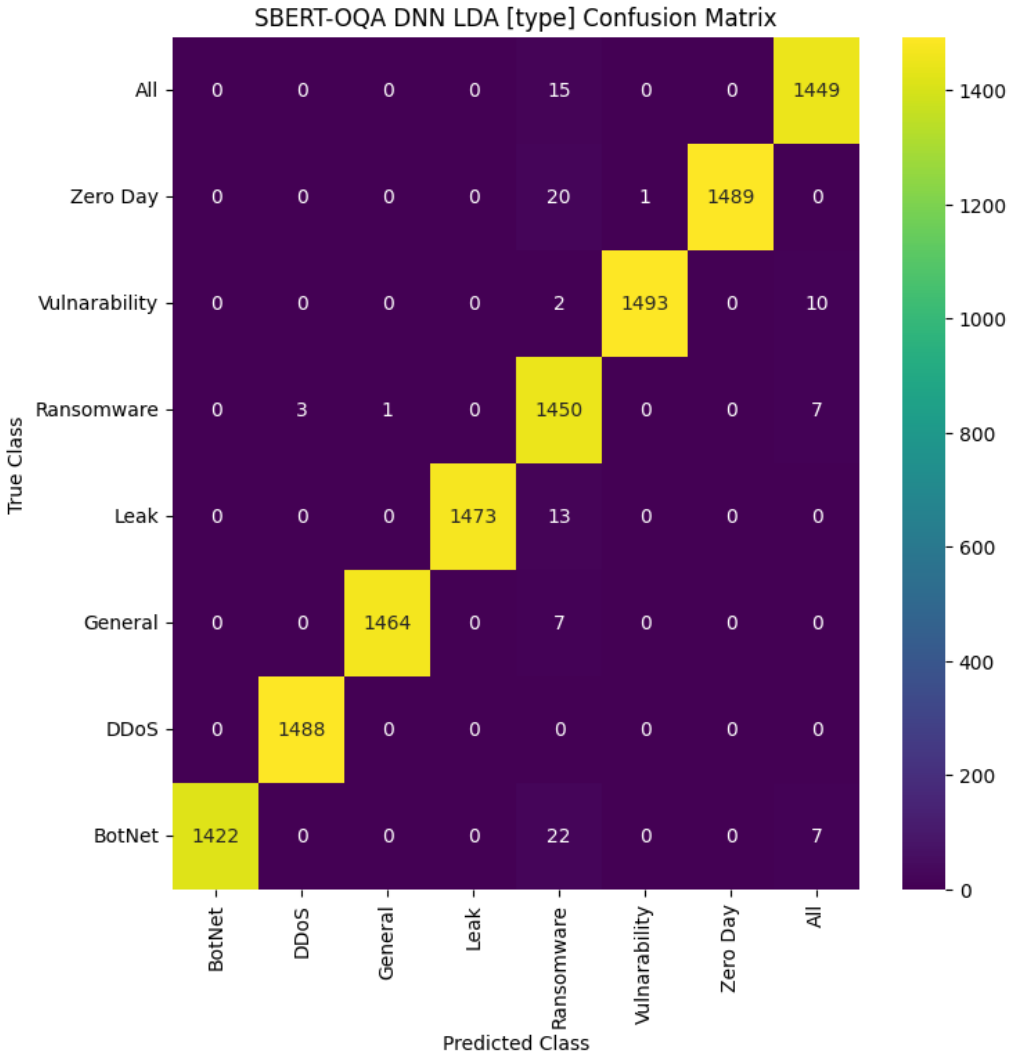

Figure 7: SBERT-OQA DNN LDA Confusion Matrix

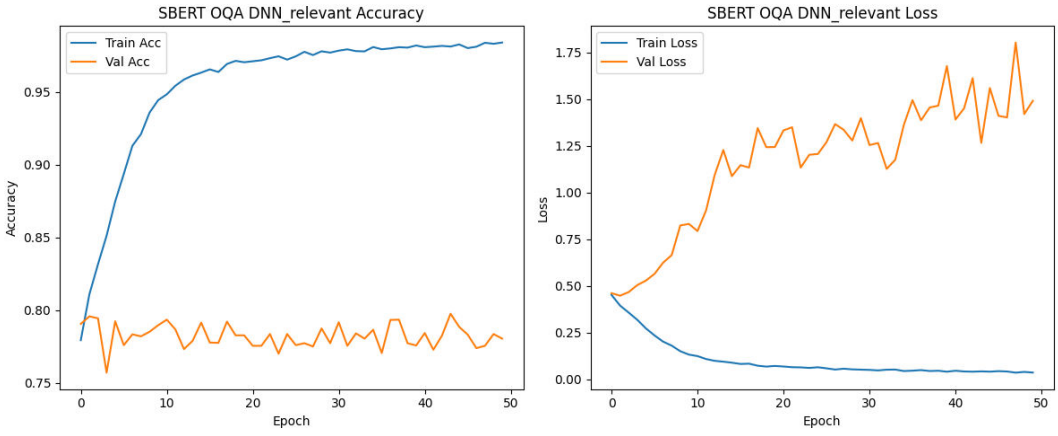

Figure 8: SBERT OQA DNN Accuracy and Loss

Figure 8 shows the training and validation performance of the SBERT-OQA DNN model for the relevant classification task across 50 epochs, illustrating accuracy and loss trends. The training accuracy increases steadily from approximately 0.78 to around 0.98, indicating effective learning and convergence of the model. In contrast, validation accuracy fluctuates within a narrower range of 0.76–0.80, stabilizing around 0.78, suggesting limited generalization improvement beyond early epochs.

Correspondingly, the training loss decreases sharply from about 0.45 to nearly 0.04, reflecting successful optimization, while the validation loss rises from approximately 0.45 to values between 1.2 and 1.8, indicating increasing divergence between training and validation loss. This pattern suggests overfitting in the standalone DNN, which justifies the subsequent use of LDA as a discriminative layer to improve generalization performance in the final SBERT-OQA DNN–LDA model.

Figure 9 shows the home page of the TweetSentinel web application developed for real-time social media threat identification. The interface prominently displays the system title "TweetSentinel For Real-Time Social Media Threat Identification" at the top, along with navigation options such as Home, Login, and Register, enabling user access control. The central section features a visual security theme with a digital lock icon, symbolizing cyber protection, accompanied by the welcome message "Welcome To TweetSentinel". This section emphasizes the system's AI-powered capability for detecting and classifying cyber threats from social media streams. Social media platform icons (Twitter, Facebook, and Instagram) are positioned on the left, indicating multi-platform data relevance. Overall, the figure highlights a user-friendly, security-focused interface designed to support efficient threat monitoring and analysis.

Figure 9: Home Page

Figure 10 shows the prediction results generated by the TweetSentinel system for unseen social media data. Each row represents an individual tweet along with its associated structured fields, including the raw text, embedded tweet metadata, Watson NLP analysis, annotation, URLs, and destination URLs. The annotation column displays predicted contextual labels such as irrelevant, business, and threat, demonstrating the system's ability to automatically classify tweet intent. For example, multiple tweets are labeled as irrelevant, while security-focused content is correctly identified as threat. The presence of extracted URLs and resolved destination links highlights the system's capability to analyze external references associated with tweets. The figure illustrates the successful integration of preprocessing, semantic feature extraction, and trained classification models to produce real-time threat identification outputs in a structured and interpretable format.

	text	tweet	watson	annotation	urls	destination_url
0	@MarkHannaCrypto Its why I unfollowed him/her ...	{'created_at': 'Thu Sep 06 13:02:07 +0000 2018...	{'usage': {'text_units': 1, 'text_characters':...	NaN	['https://twitter.com/i/web/status/10376873896...	https://twitter.com/i/web/status/1037687389626...
1	@FuddBot (cont'd) ... Spikes in tweet volume: ...	{'created_at': 'Mon Sep 10 18:39:44 +0000 2018...	{'usage': {'text_units': 1, 'text_characters':...	irrelevant	['https://twitter.com/i/web/status/10392219034...	https://twitter.com/i/web/status/1039221903469...
2	EXPERT PITCH: WVU experts available to talk ab...	{'created_at': 'Tue Sep 11 19:47:03 +0000 2018...	{'usage': {'text_units': 1, 'text_characters':...	irrelevant	['http://bit.ly/2oYEpef']	https://wvutoday.wvu.edu/media-center-blog/201...
3	Leader of DDoS-for-Hire Gang Pleads Guilty to ...	{'created_at': 'Thu Sep 06 16:16:03 +0000 2018...	{'usage': {'text_units': 1, 'text_characters':...	NaN	['https://ift.tt/2oLPsaL']	https://krebsonsecurity.com/2018/09/leader-of-...
4	It Takes an Average 38 Days to Patch a Vulnera...	{'created_at': 'Fri Sep 07 13:00:46 +0000 2018...	{'usage': {'text_units': 1, 'text_characters':...	business	['http://ow.ly/Mqs630llhXZ']	https://www.darkreading.com/cloud-security/it-...
5	@Opulxx @Jirxeh @JakeZenith @n9ire Yeah same,	{'created_at': 'Sun Sep 09 14:25:06 +0000	{'usage': {'text_units': 1, 'text_characters':...	threat	[]	NaN

Figure 10: Prediction Results

Table 1 presents the overall performance of the three models for multi-class threat type classification. The SGD classifier achieves strong results with an accuracy of 94.60%, precision of 94.49%, recall of 94.55%, and F1-score of 94.49%, indicating reliable and balanced performance. The Complement NBC model performs comparatively lower, with an accuracy of 81.58%, precision of 81.24%, recall of 81.44%, and F1-score of 80.52%, showing difficulty in handling complex and overlapping threat classes. In contrast, the DNN + LDA model delivers the best performance, achieving 99.09% accuracy, 99.11% precision, 99.09% recall, and 99.09% F1-score, demonstrating the effectiveness of deep feature extraction combined with discriminant analysis for cyber threat classification.

Table 1: Performance Comparison (Type Classification)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
SGD	94.60	94.49	94.55	94.49
Complement NBC	81.58	81.24	81.44	80.52
DNN + LDA	99.09	99.11	99.09	99.09

Table 2 reports precision values, reflecting the correctness of predicted class labels. The SGD classifier maintains high precision across most classes, such as DDoS (0.99) and Leak (0.98), but shows reduced precision for Ransomware (0.84) and All (0.88). The Complement NBC model demonstrates lower precision, especially for BotNet (0.72), Ransomware (0.72), and All (0.72), suggesting higher false-positive rates. In comparison, the DNN + LDA model achieves exceptionally high precision, reaching 1.00 for most classes and maintaining 0.95 for Ransomware and 0.98 for All, indicating highly reliable class predictions.

Table 2: Precision Comparison Across Classes

Class	SGD	Complement NBC	DNN + LDA
BotNet	0.97	0.72	1.00
DDoS	0.99	0.89	1.00
General	0.97	0.84	1.00
Leak	0.98	0.96	1.00
Ransomware	0.84	0.72	0.95
Vulnarability	0.95	0.77	1.00
Zero Day	0.97	0.88	1.00
All	0.88	0.72	0.98

Table 3 shows the F1-score comparison summarizes the balance between precision and recall for each class. The SGD classifier attains F1-scores of 0.78 for both True and False classes, reflecting stable but moderate performance. The Complement NBC model records lower F1-scores of 0.72 for True and 0.74 for False, indicating weaker overall classification balance. The DNN + LDA model significantly outperforms the other methods, achieving an F1-score of 0.94 for both classes, demonstrating its robustness and effectiveness in accurately identifying relevant and non-relevant tweets.

Table 3: F1-Score Comparison Across Classes

Class	**SGD**	**Complement NBC**	**DNN + LDA**
True	0.78	0.72	**0.94**
False	0.78	0.74	**0.94**

5. CONCLUSION

This research presented TweetSentinel, an AI-driven framework for real-time threat detection in social media streams, with a particular focus on cyber security–related content. By integrating structured feature extraction, semantic embeddings using SBERT, deep neural network–based feature learning, and classical as well as discriminative classifiers, the system effectively addressed both binary relevance detection and multi-class threat type classification. Experimental results demonstrated that the SBERT-OQA DNN + LDA model achieved the highest performance, with an overall accuracy of 99.09%, precision of 99.11%, recall of 99.09%, and F1-score of 99.09% for threat type classification. In contrast, the SGD classifier achieved 94.60% accuracy, while Complement Naive Bayes reached 81.58% accuracy, confirming the superiority of deep feature representations over purely statistical methods.

Class-wise evaluation further highlighted the robustness of the proposed approach, particularly for complex and overlapping attack categories. The DNN + LDA model achieved near-perfect recall and precision values of 0.98–1.00 across all classes, including challenging categories such as Ransomware (F1-score: 0.97) and All (F1-score: 0.99), where traditional models exhibited significant confusion. Visual analysis through word clouds, bigram distributions, POS tagging, and confusion matrices validated that the dataset captures rich cyber threat semantics and that the model effectively learns discriminative patterns from social streams. The successful deployment of the TweetSentinel web interface further demonstrates the system's practical applicability for real-time monitoring and threat intelligence generation.

REFERENCES

[1] K. Kundiya and Y. Haribhakta, "A systematic review on insider threat detection using natural language processing," International Journal of Information Security, vol. 24, no. 6, p. 227, 2025.

[2] L. Priyanka and S. Gaddam, "Intelligent NLP-driven framework for automated detection and profiling of emerging cyber threats," American Journal of AI Cyber Computing Management, vol. 5, no. 4, pp. 148–156, 2025.

[3] D. Kovalchuk, "Utilizing large language models for automated real-time cyber threat analysis," Newsletter of Cherkasy State Technological University, vol. 30, no. 1, pp. 48–58, 2025.

[4] T. Pope, K. Lemieux-Mack, V. Sharma, A. Ogedengbe, and A. Thanvi, "Social Cipher: A multimodal framework for proactive threat detection with Sentintel," in Proc. IEEE Conf. Artificial Intelligence (CAI), 2025, pp. 896–901.

[5] K. Lipianina-Honcharenko et al., "A general method for real-time detection of information threats with a Ukraine case study," Radioelectronic and Computer Systems, no. 3, pp. 202–230, 2025.

[6] J. S. Ibitoye, "Modeling threat vectors in real-time using AI-enhanced surveillance analytics across cyber, land, air, and maritime domains," Int. J. Adv. Res. Publ. Rev., vol. 2, no. 6, pp. 440–463, 2025.

[7] R. T. Potla, "AI-powered threat detection in online communities: A multi-modal deep learning approach," Journal of Computer and Communications, vol. 13, no. 2, pp. 155–171, 2025.

[8] J. Reza, M. I. Khan, and S. A. Sarna, "Proactive cyber threat detection using AI and open-source intelligence," Journal of Computer Science and Technology Studies, vol. 7, no. 5, pp. 558–576, 2025.

[9] A. Vyas, "Revolutionizing risk: The role of artificial intelligence in financial risk management, forecasting, and global implementation," 2025.

[10] A. Odunaike, "Integrating real-time financial data streams to enhance dynamic risk modeling and portfolio decision accuracy," Int. J. Comput. Appl. Technol. Res., vol. 14, no. 8, pp. 1–16, 2025.

[11] C. J. Umechukwu, "AI-powered cyber threat intelligence: Real-time detection, prediction, and response through machine learning," Multiverse Journal, vol. 1, no. 1, pp. 1–21, 2024.

[12] M. M. I. Jabed and S. Ferdous, "Integrating business process intelligence with AI for real-time threat detection in critical US industries," International Journal of Research and Applied Innovations, vol. 7, no. 1, pp. 10120–10134, 2024.

[13] K. S. Sangher, A. Singh, and H. M. Pandey, "LSTM and BERT-based transformer models for cyber threat intelligence for intent identification of social media platforms exploitation from darknet forums," International Journal of Information Technology, vol. 16, no. 8, pp. 5277–5292, 2024.

[14] M. Alhady and A. Abdulbasit, "A real-time AI-powered framework for integrated malware, phishing, and security breach detection," Surman Journal of Science and Technology, vol. 6, no. 2, pp. 601–607, 2024.

[15] N. E. Vijayan, "Real-time cyber threat detection using big data analytics: A scalable framework for immediate threat response," Journal of Engineering and Applied Sciences Technology, vol. 6, no. 12, pp. 2–5, 2024.