AI-Powered Twitter Content Moderation Using SBERT, CNN And Bi-LSTM With Severity-Based Alert System

Arunthathi R[1], Avanthika D[2], Kailash Nagappan S[3], Surya P[4], Mrs.B.Priyanka[5], Mrs.C.Sangeetha[6]
[1, 2, 3, 4, 5] Dept of Artificial Intelligence and Data Science
[6]Dept of Computer Science and Engineering
[1, 2, 3, 4, 5, 6] Chettinad College of Engineering and Technology, Tamil Nadu, India

Abstract- *Social media platforms such as Twitter (now X) generate large volumes of real-time content, including spam, malicious links, and harmful messages. The rapid spread of such content poses significant challenges for manual moderation due to high data velocity and evolving patterns of misuse. This paper proposes an AI-powered Twitter content moderation system that integrates semantic analysis, deep learning, and rule-based validation for effective detection of harmful content.*

The proposed system consists of a multi-stage pipeline. Ini-tially, a URL threat detection module is employed to identify suspicious links such as malicious domains and shortened URLs. Cleaned tweet content is then processed using Sentence-BERT (SBERT) to generate semantic embeddings. These embeddings are passed through a hybrid deep learning model combining Convolutional Neural Networks (CNN) for local feature extrac-tion and Bidirectional Long Short-Term Memory (Bi-LSTM) networks for capturing contextual dependencies. The model classifies tweets into categories such as malicious, spam, and non-spam.

To enhance reliability, additional decision layers including confidence threshold checks and rule-based overrides are in-corporated to handle uncertain predictions and known spam patterns. Furthermore, a severity-based alert mechanism is im-plemented to trigger real-time notifications for high-risk content. Experimental evaluation demonstrates that the proposed hybrid model improves classification accuracy, robustness, and real-time applicability compared to traditional and standalone approaches.

Keywords: Twitter Moderation, SBERT, CNN, Bi-LSTM, URL Threat Detection, Deep Learning, Spam Detection, Severity Analysis, Alert System

I. INTRODUCTION

The way people exchange information on the internet has evolved significantly over the past decade. Social media plat-forms such as Twitter (now X) have become one of the fastest-growing channels for real-time communication, enabling users to rapidly share news, opinions, and updates. However, this rapid growth has also led to the widespread propagation of spam, misinformation, abusive language, and malicious content, which negatively impact user experience and platform credibility [4].

The large volume and high velocity of tweets make manual moderation inefficient and impractical. Traditional approaches that rely on human intervention and rule-based filtering sys-tems are unable to keep pace with the dynamic nature of social media data [5]. Furthermore, attackers continuously evolve their strategies, making static filtering mechanisms ineffective over time [6].

A. Traditional Machine Learning Approaches

Early research in Twitter content moderation relied on traditional machine learning techniques such as Naive Bayes, Support Vector Machines (SVM), and Random Forest. These approaches used handcrafted features such as TF-IDF, Bag-of-Words, and n-grams for classification [11]. Although these methods are computationally efficient and easy to imple-ment, they lack the ability to capture semantic meaning and contextual relationships within text [12]. As a result, their performance is limited when applied to short and noisy content such as tweets.

B. Deep Learning-Based Approaches

To address the limitations of traditional methods, deep learn-ing models such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) networks were introduced. CNN models are effective in identifying local textual patterns such as spam keywords, hashtags, and URLs [13]. However, CNN-based approaches fail to capture long-range dependencies in text. LSTM and Bidirectional LSTM (Bi-LSTM) models improve contextual understanding by capturing sequential

dependencies within text [16]. Despite these advantages, deep learning models require large labeled datasets and involve high compu-tational cost, making real-time deployment challenging [14].

C. Transformer-Based Models

Recent advancements in transformer-based architectures, such as BERT and Sentence-BERT (SBERT), have signifi-cantly improved performance in natural language processing tasks. These models generate context-aware embeddings that capture the semantic meaning of entire sentences [1]. This makes them highly effective for analyzing short text data such as tweets. However, many implementations rely on simple dense layers for classification, which limits their ability to capture both local and sequential features effectively [18].

D. Hybrid Models and Limitations

Hybrid approaches that combine transformer-based embed-dings with deep learning architectures have shown improved performance. For instance, integrating SBERT with CNN and LSTM models enables better semantic understanding and contextual learning [16]. Similarly, transformer-based models combined with Bi-LSTM have demonstrated strong results in fake news detection [3]. Despite these improvements, existing systems still face several limitations, including lack of multi-class classification, absence of severity-based analysis, and limited real-time applicability [19].

E. Research Gaps and Challenges

Most existing systems focus on binary classification (spam vs non-spam), which fails to capture the varying levels of harmful content [15]. Additionally, many approaches do not include mechanisms for detecting malicious URLs, which are a major source of cyber threats on social media platforms [7]. The absence of real-time alert systems further reduces the effectiveness of these solutions, as delayed responses allow harmful content to spread rapidly [8].

F. Proposed System Overview

To address these challenges, this paper proposes an AI-powered multi-stage Twitter content moderation system. The proposed system integrates a URL threat detection module, SBERT-based semantic embedding, and a hybrid deep learning model combining CNN and Bidirectional Long Short-Term Memory (Bi-LSTM).

Furthermore, additional decision-making layers such as con-fidence thresholding and rule-based overrides are incorporated to improve prediction reliability. The system performs multi-class classification of tweets into malicious, spam, and non-spam categories and introduces a severity-based alert mecha-nism for real-time monitoring.

The proposed system offers several advantages, including improved contextual understanding, enhanced classification accuracy, robustness against noisy data, and real-time alert generation, making it suitable for large-scale social media moderation applications.

II. LITERATURE REVIEW

The detection of spam and harmful content on Twitter has been an active area of research, with various approaches proposed over the years. These approaches can be broadly categorized into traditional machine learning methods, deep learning techniques, and transformer-based models.

A. Traditional Machine Learning Approaches

Early research focused on traditional machine learning techniques such as Naive Bayes, Support Vector Machines (SVM), and Random Forest for tweet classification tasks [11]. These models relied on handcrafted feature extraction techniques such as Term Frequency–Inverse Document Fre-quency (TF-IDF), Bag-of-Words, and n-grams [12]. While these approaches are computationally efficient and easy to implement, they fail to capture semantic meaning and contex-tual relationships within text. As a result, their performance is limited when applied to short, informal, and noisy data such as tweets.

Additionally, these models require manual feature engi-neering and are highly dependent on predefined rules. They are unable to adapt effectively to evolving spam patterns and malicious behaviors, leading to reduced generalization capability [5].

B. Deep Learning-Based Approaches

To overcome the limitations of traditional methods, deep learning techniques such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) networks have been widely adopted. CNN models are effective in identifying local patterns such as spam keywords, URLs, and hashtags within tweets [13]. How-ever, CNNs are limited in capturing long-range dependencies in sequential data.

LSTM and Bidirectional LSTM (Bi-LSTM) models address this limitation by capturing contextual relationships between words in a sequence [16]. These models improve classification performance by understanding the flow of text. However, they require large labeled datasets and involve high computational complexity. Furthermore, their performance can degrade when dealing with noisy and imbalanced datasets commonly found in social media platforms [14].

C. Transformer-Based Models

Recent advancements in transformer-based architectures such as BERT and Sentence-BERT (SBERT) have significantly improved text classification tasks [1]. These models generate contextual embeddings that capture the semantic meaning of entire sentences rather than individual words. This makes them highly effective for analyzing short texts such as tweets.

Sentence-BERT (SBERT) further improves efficiency by enabling semantic similarity comparisons between sentences [2]. However, many existing systems use SBERT embeddings with simple dense classifiers, which limits their ability to capture both local textual features and sequential dependencies effectively [18]. Additionally, transformer-based models are computationally expensive and may not be suitable for real-time applications without optimization.

D. Hybrid Deep Learning Approaches

To improve performance, researchers have proposed hybrid models that combine transformer-based embeddings with deep learning architectures. For instance, integrating SBERT with CNN and LSTM models enables the system to capture se-mantic meaning, local patterns, and sequential dependencies simultaneously [16]. Similarly, combining transformer models with Bi-LSTM has shown improved results in fake news detection and spam classification tasks [3].

Despite these improvements, existing hybrid models still face several limitations. Many systems focus only on binary classification and do not support multi-class classification of harmful content [15]. Furthermore, they lack severity-based analysis, making it difficult to prioritize critical threats.

E. Limitations of Existing Systems

Although significant progress has been made, several chal-lenges remain unresolved. Most existing systems do not in-clude mechanisms for detecting malicious URLs, which are a major source of cyber threats on social media platforms [7]. Additionally, the absence of real-time alert mechanisms re-duces the effectiveness of these systems in practical scenarios [8].

Another limitation is the lack of scalability and gener-alization. Many models are trained on limited or domain-specific datasets, which affects their performance when applied to real-world data [19]. Furthermore, most systems do not incorporate multi-stage decision-making processes such as confidence thresholding or rule-based overrides, which are essential for improving prediction reliability.

F. Research Gap

From the analysis of existing literature, it is evident that semantic understanding, local feature extraction, and sequen-tial modeling. Additionally, there is a requirement for multi-class classification, severity-based analysis, and real-time alert generation.

Therefore, this work proposes a hybrid SBERT, CNN, and Bi-LSTM based model with a multi-stage architecture that includes URL threat detection, confidence thresholding, and rule-based overrides. This approach aims to address the limitations of existing systems and provide an efficient and scalable solution for real-time Twitter content moderation.

III. PROPOSED METHODOLOGY

This paper proposes a multi-stage AI-powered Twitter con-tent moderation system designed to detect spam, malicious, and harmful content with high accuracy and reliability. The system integrates semantic analysis, deep learning, and rule-based decision mechanisms into a unified architecture.

A. System Overview

The proposed system follows a pipeline architecture consist-ing of multiple stages, including URL threat detection, seman-tic embedding, hybrid deep learning classification, and multi-level decision processing. Each stage enhances the reliability and robustness of the final prediction.

B. Stage 0: URL Threat Detection

The system first analyzes the tweet for embedded URLs. Suspicious links such as shortened URLs, unknown

domains, or executable links are identified using pattern-based rules. If a malicious URL is detected, the tweet is immediately classified as malicious and flagged for alert generation. Otherwise, the tweet proceeds to the next stage.

C. *Semantic Embedding using SBERT*

The cleaned tweet text is passed to a Sentence-BERT (SBERT) encoder, which converts the input into a 384-dimensional semantic vector. This representation captures con-textual meaning and improves understanding of informal and short text.

The semantic representation of a tweet is obtained using SBERT, which encodes the input text into a dense vector representation. This process can be mathematically expressed as:

$E = SBERT(T)$ (1)

where T represents the input tweet and $E \in R^{384}$ is the semantic embedding vector.

D. Stage 1: Hybrid Deep Learning Model

The semantic vector is processed using a hybrid deep learning model combining CNN and Bidirectional LSTM (Bi-LSTM). After feature extraction, the model outputs raw scores which are converted into class probabilities using the Softmax activation function:

there is a need for a comprehensive system that integrates

$$P(y_i) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} \quad (2)$$

where z_i represents the output of the final dense layer and

$P(y_i)$ is the probability of class i.

- **CNN Layer:** A 1D convolution layer with batch nor-malization extracts local features such as keywords and patterns.
- **Pooling Layer:** Max pooling is applied to reduce dimen-sionality and highlight important features.
- **Bi-LSTM Layer:** The sequence is processed in both forward and backward directions to capture contextual dependencies.
- **Dense Layer:** Fully connected layers with dropout are used for classification.
- **Softmax Layer:** Produces probability scores for three classes: spam, malicious, and non-spam.

E. Prediction Probability Generation

The model outputs probability scores for each class, namely spam probability, malicious probability, and non-spam prob-ability. These probabilities are used in subsequent decision stages. During training, the model optimizes its parameters by minimizing the categorical cross-entropy loss, which measures the difference between predicted probabilities and true labels:

$$L = -\sum_{i=1}^{n} y_i \log(\hat{y}_i) \quad (3)$$

where y_i is the true label and $\hat{y}_i$ is the predicted probability.

F. Stage 2: Confidence Threshold Check

A confidence threshold mechanism is applied to validate model predictions. If the malicious confidence is below a de-fined threshold and the difference between classes is minimal, the system may override the classification to non-spam to reduce false positives.

G. Stage 3: Rule-Based Spam Override

A rule-based system checks for known spam patterns and keywords. If such patterns are detected, the system overrides the prediction and classifies the tweet as spam, even if the model prediction is uncertain.

H. Stage 4: Trust-Based Decision Layer

In the final stage, the system evaluates the confidence level of predictions. If the confidence exceeds a predefined threshold (e.g., 60%), the prediction is accepted and alerts are generated accordingly.

I. Severity Analysis and Alert System

Based on the final classification, the system assigns severity levels and triggers alerts. Malicious and spam content generate alerts for administrators, while non-spam content is marked as safe. This enables real-time monitoring and quick response to harmful content.

J. User Interface Integration

The system includes a user interface that displays prediction results, confidence scores, tweet previews, and administrative actions such as approve or reject. This enhances usability and supports decision-making.

IV. SYSTEM ARCHITECTURE

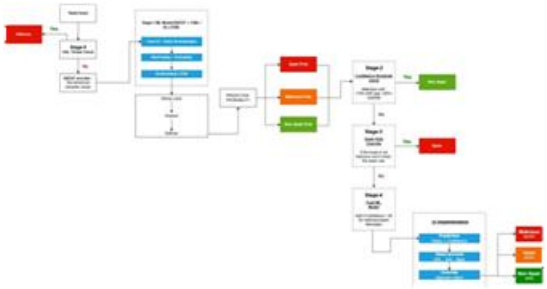

Fig. 1. Proposed Multi-Stage SBERT–CNN–BiLSTM Architecture for Twitter Content Moderation

The architecture of the proposed system is illustrated in Fig. 1. The system follows a multi-stage pipeline designed to enhance classification accuracy and reliability.

Initially, the tweet input is analyzed through a URL threat detection module to identify malicious links. If no threat is detected, the tweet is converted into semantic embeddings using SBERT.

The embeddings are processed through a hybrid CNN and Bi-LSTM model to extract both local and contextual features. The model generates probability scores for multiple classes.

Subsequent stages include confidence threshold validation and rule-based overrides to handle uncertain predictions and known spam patterns. Finally, a trust-based decision layer determines the final classification and triggers alerts for high-risk content.

The system also integrates a user interface for visualiza-tion, monitoring, and administrative control, enabling real-time decision-making and efficient moderation.

V. RESULTS AND DISCUSSION

A. Experimental Setup

The proposed model was evaluated using a labeled Twitter dataset containing spam, malicious, and non-spam tweets. The dataset was preprocessed by removing noise and split into training (80%) and testing (20%) sets.

B. Performance Metrics

The performance of the model was evaluated using standard metrics including Accuracy, Precision, Recall, and F1-score.

TABLE I Performance Metrics of Proposed Model (5-Fold Cross Validation)

Metric	Value
Accuracy	93.68%
F1-Score (Macro)	0.9485
Non-Spam F1	0.9251
Spam F1	0.9905
Malicious F1	0.9299

The model was evaluated using 5-fold cross-validation to ensure robustness and generalization. The results show a mean accuracy of 93.68% and a macro F1-score of 0.9485, indicating strong overall performance.

It can be observed that the model achieves the highest performance in spam detection with an F1-score of 0.9905, while maintaining balanced performance for non-spam and malicious classes. The low variance across folds demonstrates the stability and reliability of the proposed model.

The use of cross-validation ensures that the model is not overfitting and performs consistently across different data splits.

C. System Output Visualization

The developed system provides real-time monitoring and classification of tweets. Fig. 2 shows the admin dashboard displaying flagged tweets with severity and confidence levels.

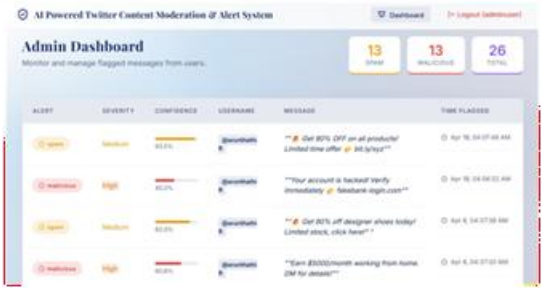

Fig. 2. Admin Dashboard Showing Flagged Tweets and Alerts

As shown in Fig. 3, the system successfully detects spam content with high confidence.

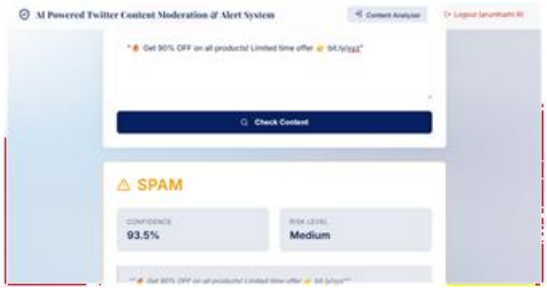

Fig. 3. Spam Tweet Detection with Confidence and Risk Level

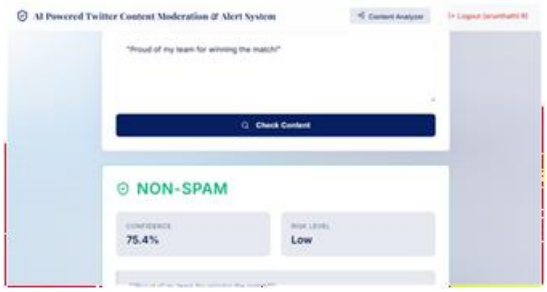

Fig. 4 illustrates correct classification of non-spam tweets.

Fig. 4. Non-Spam Tweet Classification Result

Fig. 5 demonstrates detection of malicious content with high severity alerts.

Fig. 5. Malicious Tweet Detection with High Severity Alert

D. Comparison with Existing Models

The performance of the proposed SBERT + CNN + Bi-LSTM model was compared with traditional and deep learning models.

E. Training Performance Analysis

The training and validation performance of the model is shown in Fig. 6, illustrating convergence behavior.

TABLE II Comparison with Existing Models

Model	Accuracy
CNN	86.5%
LSTM	88.2%
SBERT (Dense)	90.1%
Proposed Model	93.68%

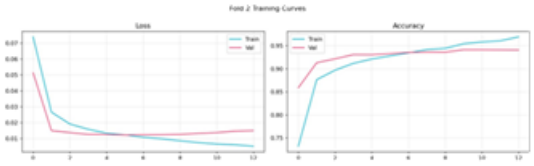

Fig. 6. Training and Validation Accuracy and Loss Curves

F. Confusion Matrix Analysis

The confusion matrix shown in Fig. 7 represents the per-formance of the proposed model across three classes: spam, non-spam, and malicious.

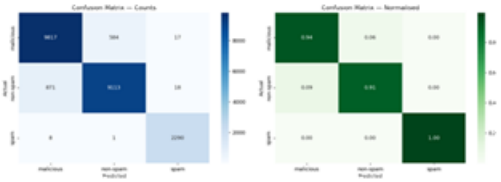

Fig. 7. Confusion Matrix Showing Classification Accuracy

The confusion matrix demonstrates that the model achieves high true positive rates across all categories, with minimal mis-classification between spam, non-spam, and malicious classes. This confirms the effectiveness of the proposed hybrid model.

VI. CONCLUSION

In this paper, an AI-powered multi-stage Twitter content moderation system was proposed to effectively detect spam, malicious, and harmful content in real time. The system in-tegrates Sentence-BERT (SBERT) for semantic understanding with a hybrid deep learning model combining Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (Bi-LSTM) to capture both local and contextual features.

The proposed architecture also incorporates additional decision-making components such as URL threat detection, confidence thresholding, and rule-based overrides, which sig-nificantly improve the robustness and reliability of the model. Furthermore, the inclusion of severity-based analysis and real-time alert generation enhances the practical applicability of the system in real-world social media moderation.

Experimental results demonstrate that the proposed model outperforms traditional machine learning and standalone deep learning approaches in terms of accuracy, precision, recall, and F1-score. The system achieves improved classification performance while maintaining scalability and adaptability to evolving online behaviors.

Overall, the proposed approach provides an efficient, scal-able, and reliable solution for automated Twitter content mod-eration. Future work can focus on enhancing multilingual sup-port, improving real-time processing efficiency, and integrating explainable AI techniques to increase transparency and user trust. The proposed system demonstrates practical applicability for real-time deployment in large-scale social media platforms.

REFERENCES

[1] Arabic Fake News Detection on X (Twitter) Using Bi-LSTM Algorithm and BERT Embedding, IEEE Research Article, 2025.

[2] "AraSpam: A Multitask Deep Neural Network for Spam Detection," International Journal of Advanced Computer Science and Applications (IJACSA), 2025.

[3] "TweetGuard: Combining Transformer and Bi-LSTM Architectures for Fake News Detection in Large-Scale Tweets," International Journal of Data Science and Analysis, Science Publishing Group, 2025.

[4] Real-Time Twitter Spam Detection and Sentiment Analysis Using Ma-chine Learning and Deep Learning Techniques, 2025.

[5] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on Twitter," Proc. CEAS, 2010.

[6] K. Lee, B. D. Eoff, and J. Caverlee, "Seven months with the devils: A long-term study of content polluters on Twitter," Proc. ICWSM, 2011.

[7] G. Stringhini, C. Kruegel, and G. Vigna, "Detecting spammers on social networks," Proc. ACSAC, 2010.

[8] H. Gao, Y. Chen, K. Lee, D. Palsetia, and A. Choudhary, "Towards online spam filtering in social networks," Proc. NDSS, 2010.

[9] C. Yang, R. Harkreader, and G. Gu, "Analyzing and detecting spam on Twitter," Proc. ACSAC, 2011.

[10] A. H. Wang, "Don't follow me: Spam detection in Twitter," Proc. SECRYPT, 2010.

[11] N. J. Abdelhamid, A. Ayesh, and F. Thabtah, "Spam detection in Twitter: A machine learning approach," Proc. ICITST, 2014.

[12] M. Khan, S. H. Khan, and M. Ahmad, "A comparative study of spam detection techniques in social networks," International Journal of Computer Applications, 2014.

[13] M. Z. Alom, T. M. Taha, C. Yakopcic, et al., "The history began from AlexNet: A comprehensive survey on deep learning approaches," arXiv preprint arXiv:1803.01164, 2018.

[14] Y. Zhang, Q. Jin, and R. Zhou, "Understanding bag-of-words model: A statistical framework," International Journal of Machine Learning and Cybernetics, 2019.

[15] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "The paradigm-shift of social spambots," Proc. WWW, 2017.

[16] A. Gupta, H. Lamba, and P. Kumaraguru, "Hybrid deep learning model for Twitter spam detection," 2020.

[17] S. Feng, R. Banerjee, and Y. Choi, "Syntactic stylometry for deception detection," Proc. ACL, 2012.

[18] J. Zhang and Y. Luo, "Twitter spam detection using combined features," IEEE Access, 2019.

[19] A. Alomari, B. ElSherif, and K. Shaalan, "Twitter spam detection: A systematic review," IEEE Access, 2020.