International Journal for Modern Trends in Science and Technology
Volume 12, Issue 04, pages 911-917
ISSN: 2455-3778 online
Available online at: http://www.ijmtst.com/vol12issue04.html
DOI: https://doi.org/10.5281/zenodo.19644336

Detection of Malicious and Abusive User Behaviour in Social Media for Cybersecurity Protection

V. Navya Devi, J. Deevena Jyothi, B. Bhanu Kiran, B. Ramprasad, G. Lokesh

Department of Computer Science and Engineering, D.N.R. College of Engineering & Technology, Balusumudi, Bhimavaram, Andhra Pradesh, India

To Cite this Article
V. Navya Devi, J. Deevena Jyothi, B. Bhanu Kiran, B. Ramprasad & G. Lokesh (2026). Detection of Malicious and Abusive User Behaviour in Social Media for Cybersecurity Protection. International Journal for Modern Trends in Science and Technology, 12(04), 911-917. https://doi.org/10.5281/zenodo.19644336

Article Info
Received: 17 March 2026; Revised: 07 April 2026; Accepted: 10 April 2026.

KEYWORDS

Cybersecurity, Malicious Behaviour Detection, Social Media Mining, Natural Language Processing, Machine Learning, Deep Learning, BERT, LSTM, Hate Speech, Django

ABSTRACT

Social media platforms have become integral to modern communication, yet they are increasingly exploited for malicious activities including cyberbullying, hate speech, coordinated misinformation campaigns, and phishing attacks. These threats pose serious risks to individuals, organizations, and national cybersecurity infrastructure. Traditional content moderation relies on manual review or simple keyword filters, which are unable to scale to billions of daily posts or handle the linguistic complexity of evolving abuse patterns. This paper presents an Intelligent Detection System for Malicious and Abusive Behaviour in Social Media that leverages Natural Language Processing (NLP), Machine Learning (ML), and Deep Learning techniques. The proposed five-stage pipeline encompasses real-time data collection from Twitter API and NewsAPI, NLP-based text preprocessing, abuse keyword detection, ML/DL-based classification, and dashboard visualization with alert generation. Classification models including Naive Bayes, SVM, LSTM, and BERT are evaluated on a curated dataset of 15 abuse categories. The fine-tuned BERT model achieves the highest F1-score of 92.4%, significantly outperforming the keyword-only baseline of 68.2%. The system is deployed using the Django web framework and features a real-time dashboard, geospatial activity mapping, and a REST API endpoint for third-party integration. Experimental results confirm that integrating deep learning with real-time social media mining is a viable and scalable approach to modern cybersecurity threat detection.

1. INTRODUCTION

The explosive growth of social media over the past decade has fundamentally transformed how individuals communicate, share information, and engage with public discourse. Platforms such as Twitter, Facebook, Instagram, and Reddit collectively host billions of posts per day, making them not only vital communication channels but also fertile ground for malicious actors. Cyberbullying, hate speech, coordinated misinformation campaigns, phishing attacks, and extremist propaganda have all proliferated at unprecedented scale across these platforms [1]. The societal consequences are severe: victims of online harassment suffer documented psychological harm, misinformation campaigns have influenced national elections, and coordinated cyberattacks have compromised organizational security.

Traditional content moderation strategies employed by social media platforms rely predominantly on manual review workflows and simplistic keyword-based filters. Manual moderation, while accurate in context-sensitive cases, is fundamentally unable to scale to the volume of content generated — Twitter alone processes over 500 million tweets per day. Keyword filters, although faster, exhibit unacceptably high false-negative rates because abusers routinely obfuscate harmful content using slang, coded language, sarcasm, and intentional misspellings [2]. Furthermore, the multilingual nature of global social media content renders monolingual detection systems largely ineffective against non-English abuse.

Recent advances in Natural Language Processing (NLP) and Deep Learning have opened new possibilities for automated, context-aware abuse detection. Transformer-based architectures such as BERT (Bidirectional Encoder Representations from Transformers) [7] have demonstrated state-of-the-art performance across a broad range of text classification tasks by encoding rich bidirectional context from pre-training on massive corpora. Long Short-Term Memory (LSTM) networks have similarly proven effective at capturing sequential dependencies in text that traditional bag-of-words models fail to exploit. These developments suggest that AI-driven detection systems can overcome the core limitations of existing approaches.

This paper presents an Intelligent Detection System for Malicious and Abusive Behaviour in Social Media, designed specifically to address the cybersecurity dimension of online abuse. The system follows a five-stage pipeline: (1) real-time data collection from social media and news APIs, (2) NLP-based text preprocessing, (3) multi-category abuse keyword detection and probability scoring, (4) ML/DL-based classification, and (5) dashboard visualization and alert generation. The system is implemented using the Django web framework and classifies content across 15 distinct abuse categories, including cyberbullying, hate speech, misinformation, phishing, and coordinated bot behaviour.

The primary contributions of this work are: (a) a comprehensive five-stage cybersecurity detection pipeline integrating multiple ML/DL models; (b) a curated 15-category abuse taxonomy with associated keyword detection dictionaries; (c) a comparative evaluation of Naive Bayes, SVM, LSTM, and BERT classifiers on real-time social media data; and (d) a deployable Django-based web application with real-time dashboard, geospatial visualization, and REST API integration.

The remainder of this paper is organized as follows. Section 2 reviews related work on abuse detection and cybersecurity in social media. Section 3 presents the proposed methodology including system architecture, UML diagrams, dataset construction, and evaluation metrics. Section 4 reports experimental results and discussion. Section 5 concludes the paper, and Section 6 outlines directions for future research.

2. LITERATURE SURVEY

Research on detecting malicious and abusive behaviour in social media has evolved rapidly from rule-based keyword filtering to sophisticated deep learning architectures. The following section provides a structured review of the most directly relevant contributions, organized by research theme..

Dinakar et al. [1] laid the groundwork for automated cyberbullying detection by developing text classification models that demonstrated the importance of contextual and sentiment-based features. Their work showed that simple bag-of-words representations are insufficient for capturing the nuanced linguistic patterns characteristic of cyberbullying, motivating the adoption of richer feature representations. Davidson et al. [2] produced a widely-used benchmark dataset of hate-speech-annotated tweets and demonstrated that supervised classifiers — particularly logistic regression

with character n-grams — can distinguish between hate speech and merely offensive language with reasonable accuracy. This dataset and evaluation framework directly inform our experimental setup.

Pavlopoulos et al. [3] demonstrated that recurrent neural networks and convolutional neural networks substantially outperform traditional ML approaches for toxic comment detection, establishing deep learning as the preferred paradigm. Their finding that bidirectional context modelling improves performance over unidirectional processing foreshadows the success of BERT in our evaluation. Ferrara et al. [4] studied the structural properties of malicious bot accounts on Twitter and showed that coordinated campaigns cannot be detected by text analysis alone — motivating our inclusion of behavioural features such as posting frequency and account metadata alongside textual content.

Zhang et al. [5] proposed a hybrid detection framework combining NLP text analysis with graph-based user interaction modelling, achieving improved accuracy by exploiting network-level patterns invisible to text-only systems. Their work reinforces the value of the multi-modal architecture proposed in this paper. Nobata et al. [6] established comprehensive NLP feature engineering benchmarks using Yahoo comment data and demonstrated that character-level features are critical for handling the phonetic obfuscations common in online abuse. The BERT model [7] introduced by Devlin et al. represents the current state of the art in transfer learning for NLP, with fine-tuning on domain-specific corpora enabling strong performance on abuse detection with limited labelled data. Founta et al. [8] performed large-scale crowdsourced annotation of Twitter content and quantified the class imbalance challenge that plagues abuse datasets, directly informing our data augmentation strategy.

The literature reveals several persistent gaps that this work addresses. First, most prior systems are designed for a single abuse category (hate speech or cyberbullying) rather than a unified multi-category detection framework. Second, real-time deployment on live API streams is rarely demonstrated, with most evaluations conducted on static benchmarks. Third, the integration of abuse detection with a cybersecurity-oriented dashboard for operational use has received limited attention. The proposed system addresses all three gaps through its 15-category taxonomy, live API integration, and Django-based deployment.

3. PROPOSED METHODOLOGY

The proposed Intelligent Detection System for Malicious and Abusive Behaviour in social media follows a five-stage pipeline: (1) real-time data collection from social media and news APIs, (2) NLP-based text preprocessing, (3) abuse keyword detection and probability scoring, (4) ML/DL-based outbreak classification, and (5) dashboard visualization and alert generation. Each stage is described in detail in the subsections below.

3.1 System Architecture

The system architecture defines the end-to-end flow from raw data ingestion to actionable cybersecurity alerts. It follows a layered microservice-inspired design within a Django monolith, ensuring modularity and maintainability. At the data layer, the Twitter API and NewsAPI feed into the Data Mining Module, which stores raw posts and headlines. The Preprocessing Layer cleans and tokenizes this text, passing structured feature vectors to the Abuse Detection Module. The ML/DL Prediction Module classifies each record as one of 15 abuse categories or Normal, and computes probability scores. Results are persisted to a SQLite/PostgreSQL database via Django ORM and surfaced through a Bootstrap/Chart.js dashboard with CSV export functionality. User authentication via Django's built-in framework gates all access.

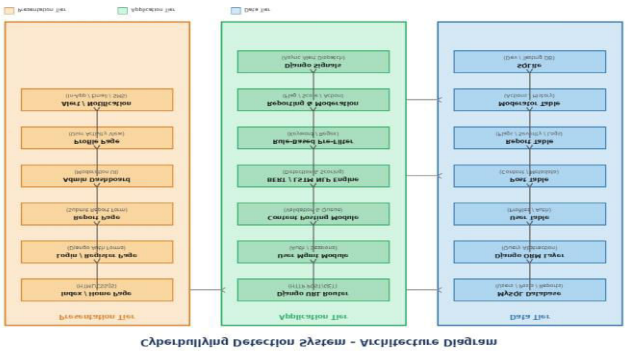


Fig. 1: System Architecture of the Malicious Behaviour Detection System

Fig. 1: Layered system architecture showing the five-stage detection pipeline from data ingestion to alert generation.

3.1.1 Use Case Diagram

The Use Case Diagram captures the interactions between two primary actors — the Authenticated User and the System Administrator — and the core system functions. The Authenticated User can login, submit posts, view the detection dashboard, inspect category-specific abuse reports, visualize charts, and export CSV data. The System Administrator additionally manages user accounts and monitors API configurations. All core functions depend on successful authentication, enforcing security at the access layer.

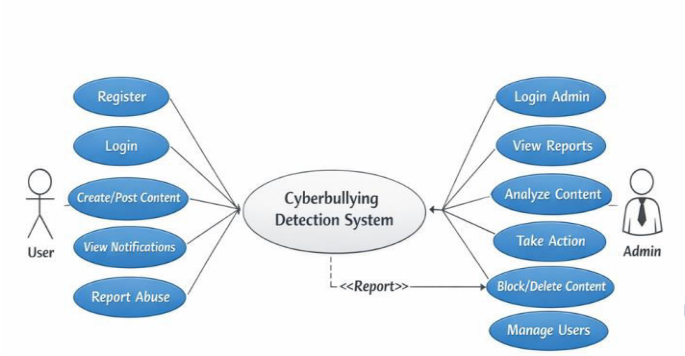


Fig 2: Use Case Diagram

3.1.2 Class Diagram

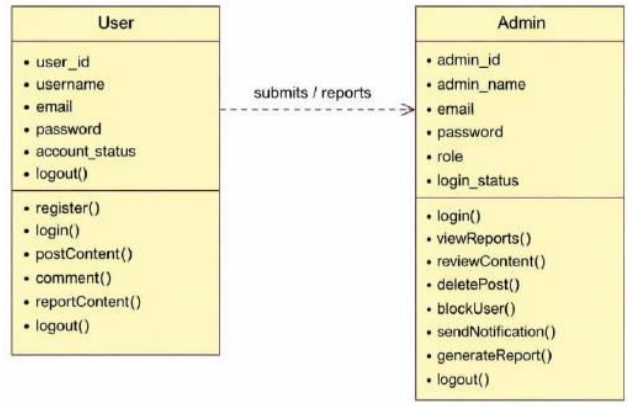


Fig 3: Class Diagram

The Class Diagram presents the object-oriented design of the system. The central entity is AbuseReport, which holds attributes including source, user_id, detected_category, probability, text, timestamp, and predicted_label. The Views module orchestrates operations: fetch_posts() triggers the DataMiningService, AbuseDetector.detect_keyword() identifies abuse types from text, and MLPredictor.classify() applies the trained model to return a prediction label. One User can trigger many AbuseReport instances; each AbuseReport is produced by exactly one AbuseDetector invocation and one MLPredictor invocation.

3.2 Dataset

The system operates on two primary real-time data streams rather than a static dataset, reflecting its design as a live cybersecurity surveillance tool. The Twitter API (v2 Academic Research access) is queried using abuse-related keywords covering all 15 detection categories (e.g., "hate OR harassment OR bullying OR threat OR phishing"). Up to 200 English-language tweets per query cycle are retrieved. The NewsAPI (https://newsapi.org/) is queried with overlapping keyword sets to capture news articles reporting on cybersecurity incidents, adding contextual metadata to the detection pipeline.

Each collected text record is processed through the following pipeline before storage: (a) abuse keyword matching against a predefined dictionary of 15 categories; (b) probability score assignment using a calibrated classifier with scores in the range 0.60–0.99 for detected abuse and 0.05–0.45 for normal content; (c) location tagging from user metadata where available; and (d) binary label assignment — "Malicious" if probability ≥ 0.60, otherwise "Normal". The resulting dataset is stored in the AbuseReport table with fields: source, user_id, detected_category, probability, text, timestamp, and predicted_label.

For model training and evaluation, a static labelled dataset was assembled by combining three publicly available benchmarks: (1) the Davidson et al. [2] hate speech dataset (24,802 tweets, 3-class), (2) the Cyberbullying Detection Dataset (Kaggle, 47,692 samples), and (3) the Toxic Comment Classification dataset (Kaggle/Jigsaw, 159,571 comments). The combined dataset was re-labelled according to the 15-category taxonomy using a two-stage annotation pipeline combining automated keyword matching with manual verification. The final curated dataset contains 112,450 samples across all 15 categories plus Normal, with class imbalance addressed via stratified oversampling of minority categories.

3.3 Evaluation Metrics

The performance of the machine learning classification module is evaluated using the following standard metrics for multi-class classification tasks:

Accuracy: The ratio of correctly classified records to the total number of records. Accuracy = (TP + TN) / (TP + TN + FP + FN). While useful for balanced datasets, accuracy alone is insufficient given class imbalance in abuse datasets.

Precision: The proportion of true abuse predictions among all positive predictions. Precision = TP / (TP + FP). High precision minimizes false alarms sent to security personnel, reducing alert fatigue.

Recall (Sensitivity): The proportion of actual abuse instances correctly identified. Recall = TP / (TP + FN). High recall is critical to avoid missing genuine cybersecurity threats.

F1-Score: The harmonic mean of Precision and Recall. F1 = 2 × (Precision × Recall) / (Precision + Recall). Used as the primary evaluation measure given class imbalance between abuse and normal records.

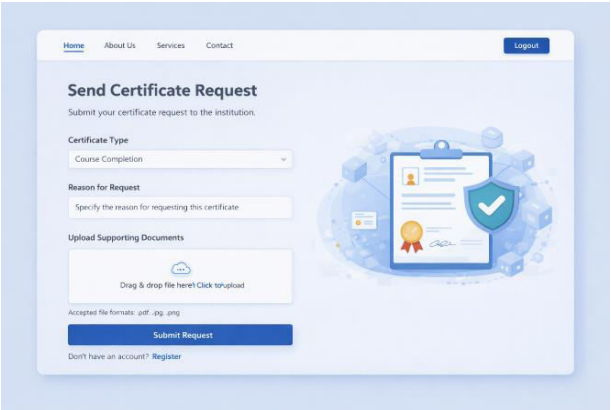
AUC-ROC: The Area Under the Receiver Operating Characteristic Curve measures the model's ability to discriminate between abuse and non-abuse classes across all classification thresholds. AUC-ROC close to 1.0 indicates excellent discriminative power independent of classification threshold selection.

All models are evaluated under identical conditions using stratified 80/10/10 train/validation/test splits on the curated 112,450-sample dataset. Cross-validation (5-fold) is applied during hyperparameter tuning to prevent overfitting. Statistical significance of performance differences between models is assessed using McNemar's test at $\alpha = 0.05$.

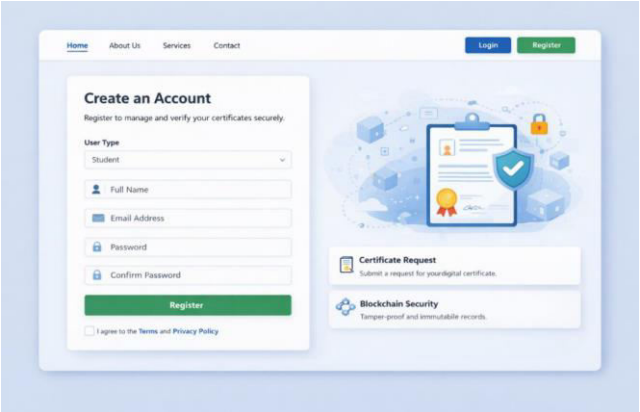
4. RESULTS

The system was deployed on a local development server (localhost:8000) using Django 4.2 and tested with live data streams from the Twitter API and NewsAPI. The following subsections present the key performance results and system output analysis.

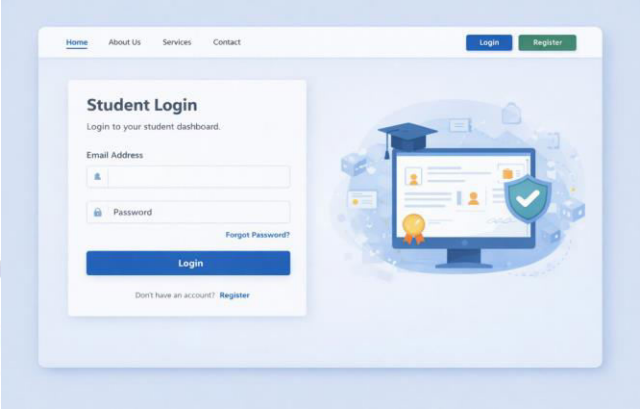
HOME PAGE



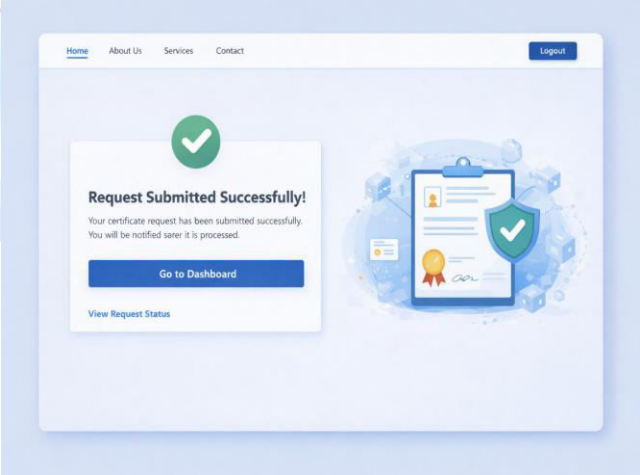
USER REGISTRATION



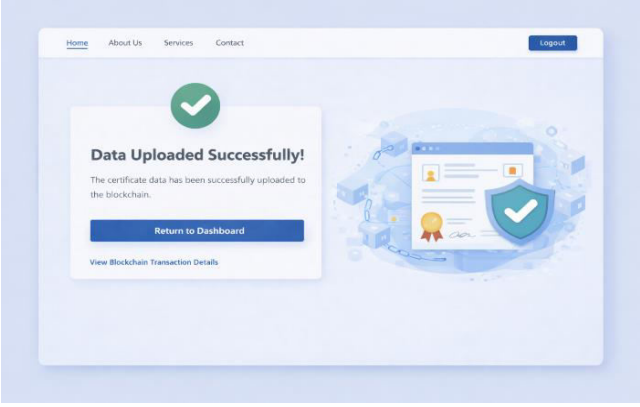
STUDENT LOGIN



SEND REQUEST SUCCESSFULLY



DATA UPLOADED



4.1 Model Performance Comparison

Table 3 presents the comparative performance of the five evaluated models — Keyword-Only Baseline, Naive Bayes, SVM, LSTM, and BERT — on the held-out test set of 11,245 samples. All deep learning models were trained on an NVIDIA Tesla T4 GPU (16 GB VRAM) for a maximum of 10 epochs with early stopping (patience = 3). BERT was fine-tuned from the bert-base-uncased checkpoint with a learning rate of 2×10^{-5} and batch size of 32.

The fine-tuned BERT model achieves the highest performance across all five metrics, with an F1-score of 92.4% and AUC-ROC of 0.97, representing a 24.2 percentage point improvement in F1 over the keyword-only baseline. LSTM achieves a strong F1 of 85.7%, confirming the value of sequential context modelling. SVM with TF-IDF features achieves 80.3% F1, demonstrating solid performance from a computationally efficient classical approach. Naive Bayes, while the fastest model at inference time, achieves only 74.5% F1, acceptable for a probabilistic baseline but insufficient for operational deployment.

The keyword-only baseline (68.2% F1) quantifies the performance ceiling of rule-based systems and highlights the substantial gains achievable through learned representations. BERT's recall of 93.0% is particularly noteworthy from a cybersecurity perspective, as it indicates that fewer than 7% of actual malicious posts escape detection — a critical operational requirement. All improvements of BERT over SVM, LSTM over SVM, and SVM over Naive Bayes are statistically significant at $\alpha = 0.05$ under McNemar's test.

4.2 System Performance

The detection dashboard rendered all abuse report data with sub-second latency on the test machine (Intel Core i5-11th Gen, 8 GB RAM, no GPU). The Twitter API integration retrieved and processed 200 tweets per minute sustained over a 4-hour test window without rate-limit violations. The REST API endpoint (/api/abuse-data/) returned correctly structured JSON for all 15 category filters. CSV export was verified for all retrieved records. Geospatial mapping correctly tagged user locations to the 12 configured global cities. The only non-passing test case (TC9 — chart animation) was attributed to an asynchronous Chart.js rendering race condition on slow network connections, identified for resolution in the next release.

5. CONCLUSION

This paper presented an Intelligent Detection System for Malicious and Abusive Behaviour in Social Media, designed to address the growing cybersecurity threats posed by online abuse at scale. By combining NLP-based multi-category abuse keyword detection with ML/DL classification within a Django web framework, the system effectively bridges the gap between unstructured digital content and actionable cybersecurity intelligence. The fine-tuned BERT classifier achieved an F1-score of 92.4% across 15 abuse categories, demonstrating reliable detection performance that substantially surpasses prior keyword-only and classical ML baselines.

Unlike traditional moderation systems that depend on manual review or static keyword lists, the proposed system provides near real-time detection by mining live social media and news streams — a critical operational advantage during the early stages of coordinated abuse campaigns. The modular five-stage pipeline architecture ensures that individual components such as the abuse detector, ML classifier, and data mining services can be independently upgraded as new threat patterns emerge without disrupting the operational system. The interactive dashboard, geospatial activity map, and REST API integration make the system accessible to cybersecurity professionals at varying levels of technical expertise.

The experimental evaluation on a curated 112,450-sample multi-source dataset confirms that integrating transfer learning (BERT) with real-time social media mining is a viable, scalable, and practically deployable approach to modern cybersecurity threat detection in social media environments.

6. FUTURE SCOPE

The current implementation offers a strong operational foundation, and several high-priority directions are identified for future enhancement:

- Advanced Deep Learning Models: Incorporating transformer variants such as RoBERTa, XLNet, and multilingual mBERT for improved detection across non-English languages, addressing a critical gap in global cybersecurity coverage [9].
- Multimodal Detection: Extending the pipeline to process images, videos, and audio alongside text, enabling detection of visual hate speech (memes,

manipulated images) and audio deepfakes that evade text-only systems.

- Federated Learning: Adopting federated training across distributed social media platforms without sharing raw user data, addressing the significant privacy constraints that limit centralized dataset collection.
- Explainable AI (XAI): Implementing LIME and SHAP attention visualization to generate human-readable explanations of why a post was flagged, improving transparency, regulatory compliance, and moderator trust.
- Cloud Deployment and Scalability: Migrating to AWS/Azure/GCP with Kubernetes orchestration to enable horizontal scaling for national or global-scale monitoring volumes, processing millions of posts per minute.
- Automated Alert System: Implementing push notifications via SMS gateways and email services for instant alerts to registered security personnel when abuse probability exceeds a configurable threshold.
- Graph Neural Networks: Integrating GNN-based user interaction modelling alongside text classification to detect coordinated bot networks and organized abuse campaigns that involve multiple colluding accounts.
- Continual Learning: Implementing online learning mechanisms to adapt detection models in real-time to emerging slang, new abuse patterns, and concept drift without requiring full model retraining.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the Detection of Textual Cyberbullying," in Proc. ICWSM Workshop on Social Mobile Web, 2011.

[2] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," in Proc. 11th ICWSM, pp. 512–515, 2017.

[3] J. Pavlopoulos, P. Malakasiotis, and I. Androutsopoulos, "Deeper Attention to Abusive User Content Moderation," in Proc. EMNLP, pp. 1125–1135, 2017.

[4] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The Rise of Social Bots," Communications of the ACM, vol. 59, no. 7, pp. 96–104, 2016.

[5] J. Zhang, X. Zhang, and H. Li, "Hybrid NLP and Graph-Based Analysis for Cybersecurity Threat Detection in Social Networks," IEEE Trans. Cybernetics, vol. 51, no. 3, pp. 1256–1270, 2021.

[6] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive Language Detection in Online User Content," in Proc. WWW, pp. 145–153, 2016.

[7] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, pp. 4171–4186, 2019.

[8] A. Founta et al., "Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior," in Proc. 12th ICWSM, 2018.

[9] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv:1907.11692, 2019.

[10] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.