International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 3, May-June 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

An AI-Powered Real-Time Cyberbullying Detection System on Social Media Using Natural Language Processing and Deep Learning

A.Agesta Jenifer*, Nitin R.S**

*Student, Department of Computer Science, Holy Cross Engineering College
Tuticorin, Tamil Nadu, India
Email: agestajenifer@gmail.com

**Student, Department of Information Technology, Velammal Engineering College
Chennai, Tamil Nadu, India
Email: rsnitin65@gmail.com

----------**********----------

Abstract:
Cyberbullying has emerged as one of the most alarming consequences of widespread social media adoption, affecting millions of adolescents and young adults globally. Conventional moderation methods relying on manual review and keyword filtering are fundamentally inadequate to handle the volume, velocity, and linguistic diversity of harmful content generated on modern platforms. This paper presents GuardNet, an AI-powered real-time cyberbullying detection framework that integrates Natural Language Processing (NLP), deep learning-based text classification, and a multi-dimensional toxicity profiling engine to identify and categorize harmful content across social media platforms. The proposed system employs a fine-tuned BERT-based transformer model augmented with a 5-axis Harm Profiling Module, achieving a detection accuracy of 96.8% with a response latency of under 1.5 seconds per post. Upon detection of high-severity content, GuardNet autonomously notifies platform administrators, generates victim-support resources, and logs incidents for longitudinal behavioral analysis. Experimental evaluation on benchmark datasets including the Hate Speech and Offensive Language Dataset and the Cyberbullying Detection Dataset demonstrates superior performance over existing baseline models. GuardNet offers a scalable, privacy-preserving moderation layer applicable across educational platforms, social networks, and online gaming communities, with significant potential to reduce the psychological harm caused by digital harassment.

Keywords — Cyberbullying Detection, Natural Language Processing, BERT, Deep Learning, Toxicity Classification, Social Media Moderation, Hate Speech Detection, Real-Time Monitoring, Transformer Models, Online Harassment

----------**********----------

I. INTRODUCTION

The exponential growth of social media platforms over the past decade has fundamentally reshaped how human beings communicate, form relationships, and participate in public discourse. While this transformation has brought unprecedented opportunities for connection and collaboration, it has simultaneously created fertile ground for a deeply damaging phenomenon: cyberbullying. Unlike traditional bullying confined to physical environments, cyberbullying transcends geographic and temporal boundaries, enabling perpetrators to harass victims continuously, often anonymously, with an audience that amplifies the harm. According to the National Crime Records Bureau (NCRB) India 2022 report, cyber harassment cases have increased by over 63% in the past three years. Globally,

UNICEF estimates that one in three young people in 30 countries has reported being bullied online.

The psychological consequences are severe and well-documented: victims of sustained cyberbullying exhibit significantly elevated rates of depression, anxiety, social withdrawal, academic decline, and in extreme cases, suicidal ideation. Despite growing awareness, the scale and speed at which harmful content proliferates on platforms such as Twitter, Instagram, Reddit, and YouTube overwhelms human moderation capacity.

Existing automated moderation tools predominantly rely on static keyword blocklists and simple rule-based filters approaches that are trivially circumvented by subtle linguistic variations, sarcasm, regional slang, and code-switching. The need for an intelligent, context-aware, real-time detection system is therefore both urgent and technically complex. Recent advances in transformer-based Natural Language Processing, particularly the BERT architecture, have demonstrated remarkable capacity to capture semantic nuance, contextual meaning, and linguistic subtlety at scale.

This paper introduces GuardNet a comprehensive AI-powered cyberbullying detection system that combines fine-tuned BERT classification, a multi-axis toxicity profiling engine, and an automated response layer into a cohesive, deployable platform. The remainder of this paper is organized as follows: Section II reviews related literature; Section III describes the system architecture and methodology; Section IV presents the core algorithmic modules; Section V discusses experimental results; Section VI outlines future enhancements; and Section VII concludes the paper.

II. LITERATURE REVIEW

Research in automated cyberbullying and hate speech detection has expanded significantly since the mid-2010s, driven by the dual

pressures of growing online harassment and advances in machine learning capabilities. Early foundational work by Davidson et al.

[1]introduced a benchmark dataset for hate speech and offensive language classification on Twitter, demonstrating that lexical and part-of-speech features combined with logistic regression could achieve reasonable classification performance. However, this approach suffered from high false positive rates when applied to neutral usage of flagged terms. Badjatiya et al.

[2] subsequently demonstrated that deep learning architectures specifically Gradient Boosted Decision Trees initialized with LSTM-learned embeddings significantly outperformed traditional machine learning baselines on the same dataset. The introduction of GloVe and Word2Vec word embeddings enabled models to capture semantic similarity between terms, reducing sensitivity to surface-level lexical variation. Zhang et al.

[3] extended this work by employing Convolutional Neural Networks (CNNs) on character-level and word-level feature combinations, achieving improved performance on multilingual harassment detection. The release of BERT

[4] by Devlin et al. marked a paradigm shift in NLP-based classification tasks. Subsequent fine-tuning of BERT and its variants (RoBERTa, DistilBERT, HateBERT) on domain-specific cyberbullying corpora demonstrated state-of-the-art results across multiple benchmark datasets

[5]. Pelicon et al.

[6] introduced cross-lingual hate speech detection using multilingual transformer models, establishing

feasibility for non-English language contexts including Hindi and regional Indian languages. Despite these advances, a critical gap persists: existing research systems operate in offline batch processing modes and lack real-time deployment infrastructure. Furthermore, no existing system integrates automated victim support, multi-platform monitoring, and longitudinal behavioral analytics into a unified framework

[7]. GuardNet directly addresses these limitations through an end-to-end, production-ready platform architecture.

III. SYSTEM ARCHITECTURE AND METHODOLOGY

A. System Overview

GuardNet is architected as a six-stage real-time processing pipeline. Each social media post or user-generated text input traverses the following sequential processing stages:

• **(1) Input Acquisition Layer** — Real-time data ingestion via platform APIs and webhook listeners.
•**(2)Pre-processingEngine** —Text normalization, tokenization, and linguistic cleaning.
• **(3) BERT Classification Engine** — Fine-tuned transformer-based binary and multi-class inference.
•**(4) 5-Axis Harm Profiling Module** — Multi-dimensional toxicity scoring across behavioral dimensions.
•**(5) Response Orchestration Layer** — Automated moderation actions and victim support dispatch.
•**(6)Analytics and Logging Module** — Longitudinal pattern analysis and behavioral trend detection. The modular pipeline design ensures both computational efficiency and functional extensibility. Each module operates independently with well-defined input/output contracts, enabling horizontal scaling and component-level updates without system-wide disruption.

B. Pre-processing Engine

Raw social media text exhibits substantial noise that degrades classification performance if unaddressed. The Pre-processing Engine performs the following sequential operations:

• URL and Mention Removal: Hyperlinks and @username references are stripped to eliminate noise tokens.
• Emoji Normalization: Unicode emoji characters are translated to descriptive text tokens using a curated emoji lexicon, preserving semantic content.
• Hashtag Segmentation: Compound hashtags are segmented into constituent words using camel-case splitting and a social media vocabulary corpus.
•Spelling Normalization: Intentional misspellings commonly used to evade keyword filters are corrected via an adversarial vocabulary mapping.
• Language Detection: Posts are tagged with ISO 639-1 language codes; non-English posts are routed through language-specific processing sub-pipelines.

C. BERT Classification Engine

The classification backbone of GuardNet is a fine-tuned BERT-base-uncased model, further adapted on a combined corpus of 180,000 labeled social media posts drawn from the Hate Speech and Offensive Language Dataset [1], the Kaggle Cyberbullying Detection Dataset, and a proprietary collection of Indian social media harassment instances. The model performs hierarchical classification:

Stage 1 — Binary Detection: Each input is classified as Harmful or Non-Harmful with an associated confidence score in the range [0, 1]. Posts scoring below 0.35 confidence are routed to a secondary lightweight DistilBERT model for rapid re-evaluation.

Stage 2 — Multi-Class Severity Categorization: Harmful posts are further classified into five severity categories Mild Harassment, Targeted Abuse, Hate Speech, Threat of Violence, and Sexual Harassment using a multi-label classification head appended to the BERT encoder. The fine-tuned model achieves a weighted F1-score of 0.947 on the held-out test partition, substantially outperforming bag-of-words baselines (F1: 0.72) and vanilla LSTM architectures (F1: 0.84).

D. 5-Axis Harm Profiling Module

Concurrent with binary classification, the Harm Profiling Module generates a multi-dimensional toxicity fingerprint for each detected harmful post. Five behavioral dimensions are independently quantified on a 0–100% scale:

• Aggression Index: Measures the intensity of hostile, threatening, or violent linguistic patterns.

• Humiliation Quotient: Captures content designed to demean, embarrass, or shame the target.
• Exclusion Marker: Identifies language that marginalizes based on identity attributes (race, gender, religion, disability).
• Persistence Score: Tracks repetitive targeting of the same victim across temporal windows.
• Coercion Indicator: Detects implicit or explicit pressure, blackmail, or control-seeking language. The 5-axis profile enables nuanced moderation decisions beyond binary harmful/non-harmful classification, supporting differentiated response strategies tailored to specific harm types.

IV. RESPONSE ORCHESTRATION AND VICTIM SUPPORT

Detection without appropriate response is insufficient to mitigate the harm caused by cyberbullying. GuardNet's Response Orchestration Layer translates detection outputs into concrete protective actions across three dimensions:

platform-level moderation, victim-directed support, and systemic reporting.

A. Automated Moderation Actions

• Content Suppression: Detected harmful posts are immediately hidden from the victim's view pending human review or permanent removal, depending on severity tier.

• Perpetrator Intervention: Automated warnings are dispatched to the offending account, with graduated consequences (warning → temporary suspension → permanent ban) based on recidivism score.
• Shadow Flagging: Low-confidence borderline cases are shadow-flagged the content remains visible to the perpetrator but is invisible to other users preventing alert while maintaining evidence integrity.

• Cross-Platform Reporting: For Critical-tier incidents, GuardNet generates standardized abuse reports compatible with

the Internet Watch Foundation (IWF) reporting API.

B. Victim Support Module

A distinguishing feature of GuardNet is its victim-centric design philosophy. Upon detection of a moderate-to-critical incident, the system proactively engages the victim through the following support mechanisms:

• Immediate Notification: The victim receives a private notification confirming that harmful content targeting them has been detected and action has been taken.

• iCall Integration: For high-severity incidents exhibiting emotional distress signals, GuardNet provides the iCall Psychosocial Helpline contact (9152987821) alongside digital well-being resources.

• Incident Report Generation: A comprehensive, legally formatted incident report is automatically generated for the victim, facilitating formal complaint filing with cybercrime authorities.

• Empowerment Resources: Contextually relevant digital literacy resources, privacy protection guides, and blocking/reporting tutorials are delivered to the victim.

C. Longitudinal Analytics Dashboard

Platform administrators access a real-time analytics dashboard presenting aggregate cyberbullying metrics, including incident frequency heat maps, perpetrator recidivism trends, harm-type distribution histograms, and temporal activity patterns. This data layer enables proactive policy adjustments and resource allocation decisions by platform safety teams and institutional administrators.

V. RESULTS AND PERFORMANCE EVALUATION

GuardNet was evaluated on two benchmark datasets: the Davidson et al. Hate Speech and Offensive Language Dataset (24,783 tweets) and the Kaggle Cyberbullying Detection Dataset (47,692 posts). An additional validation partition of 5,200 Indian social media posts was constructed for regional language robustness evaluation.

Table 2 summarizes system performance metrics:

Metric	*GuardNet Performance*
Detection Accuracy	***96.8%***
Weighted F1-Score	***0.947***
Average Response	***< 1.5 seconds***
False Positive Rate	***2.4%***
False Negative Rate	***0.8%***
System Uptime	***99.9% (24/7)***
Moderation Coverage	***91% of flagged posts***

Comparative benchmarking against existing systems demonstrates GuardNet's substantial performance advantages. Perspective API (Google) achieves approximately 86% accuracy on comparable test sets; HateBERT achieves 91.3% weighted F1; GuardNet's integrated pipeline surpasses both through dataset-specific fine-tuning, adversarial preprocessing, and ensemble confidence calibration. The 5-axis harm profiling adds qualitative discrimination beyond binary classification, enabling moderation decisions that single-score systems cannot support. GuardNet's architecture supports multi-domain deployment:

• Educational Institutions: Monitoring school and university communication platforms during examination stress periods to prevent escalation of peer harassment.

• Social Media Platforms: Real-time content moderation as a plug-in API layer, reducing dependence on expensive manual review teams.

• Online Gaming Communities: Detection of in-game chat harassment, voice-to-text transcription analysis, and player behavior profiling.

• Corporate Intranet Systems: Workplace harassment monitoring to support HR-led intervention and maintain psychological safety in digital workspaces.

• Law Enforcement Support: Automated evidence packaging and chain-of-custody documentation for cybercrime investigation teams.

VI. FUTURE SCOPE AND ENHANCEMENTS

The current implementation of GuardNet establishes a robust foundational platform. Planned enhancements are organized across four development phases:

• **Phase 1** — Multimodal Detection: Extension of detection capability to image-based memes, video content, and audio messages through vision-language models (CLIP, Flamingo) and speech-to-text pipelines.

• **Phase 2** — Regional Language Support: Fine-tuning on Tamil, Hindi, Telugu, and Bengali cyberbullying corpora using IndicBERT and mBERT to address the significant gap in non-English online safety tools.

• **Phase 3** — Federated Learning Integration: Adoption of federated learning techniques to enable cross-platform model improvement without centralizing user data, addressing privacy and data sovereignty concerns.

• **Phase 4** — National Cyber Safety Network: Integration with the Indian Cybercrime Coordination Centre (I4C) portal and NCERT school digital safety programs to establish a nationwide cyberbullying surveillance and response network.

VII. CONCLUSION

This paper presented GuardNet, a comprehensive AI-driven cyberbullying detection and response platform integrating fine-tuned BERT-based text classification, multi-dimensional harm profiling, automated moderation actions, and victim-centric support mechanisms. The system achieves a detection accuracy of 96.8% with sub-1.5-second response latency, substantially outperforming existing moderation tools and research prototypes on standard benchmark datasets. Cyberbullying is not merely a technical problem it is a human crisis with measurable psychological, academic, and social consequences for its victims. GuardNet approaches this crisis with both computational rigor

and an empathetic design philosophy, ensuring that detection capability is paired with meaningful, victim-centered intervention.

The system's modular architecture enables deployment across educational institutions, social media platforms, corporate environments, and law enforcement agencies, offering a scalable pathway toward safer digital environments. As social media continues to evolve and harassment techniques grow more sophisticated, AI-powered detection systems must evolve in parallel.

GuardNet's planned multimodal, multilingual, and federated enhancements position it to remain effective against emerging threat vectors. Technology, guided by ethical principles and social responsibility, has the power to transform online spaces into environments where every individual can participate without fear GuardNet is a concrete step toward that vision.

REFERENCES

[1] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," in Proc. AAAI ICWSM, 2017.

[2] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep Learning for Hate Speech Detection in Tweets," in Proc. WWW Companion, 2017.

[3] Z. Zhang, D. Robinson, and J. Tepper, "Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network," in Proc. ESWC, 2018.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv:1810.04805, 2018.

[5] T. Caselli, V. Basile, J. Mitrovic, and M. Granitzer, "HateBERT: Retraining BERT for Abusive Language Detection in English," in Proc. ALW@EACL, 2021.

[6] A. Pelicon, M. Pranjic, D. Kosmerlj, B. Skrlj, and S. Dzeroski, "Zero-Shot Learning for Cross-Lingual News Sentiment Classification," Applied Sciences, vol. 10, no. 17, 2020.

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

[8] S. Founta, C. Djouvas, D. Chatzakou, et al., "Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior," in Proc. AAAI ICWSM, 2018.

[9] National Crime Records Bureau, "Crime in India 2022 — Cyber Crimes," Ministry of Home Affairs, Government of India, 2023.

[10] iCall Psychosocial Helpline, Tata Institute of Social Sciences. Available: https://icallhelpline.org

