

Multimodal Cyberbullying Detection and Neutralization System in Social Media using Generative AI

Santhi¹, Preethiga², Prema³, Sudarshan⁴

¹Professor, Dept. of Information Technology, Puducherry Technological University, Puducherry, India

^{2,3,4}Under Graduate Students, Dept. of Information Technology, Puducherry Technological University, Puducherry, India

Abstract - The exponential growth of social media platforms has precipitated a rise in cyberbullying, presenting severe psychological and emotional risks that traditional reactive, text-based detection mechanisms fail to mitigate. Modern harassment is increasingly characterized by contextual complexity, sarcasm, and multimodal integration, where abusive content is often embedded within images, emojis, or memes to bypass standard filters. To address these challenges, this paper proposes a Multimodal Cyberbullying Detection and Generative Semantic Neutralization System (MCDNS-SMGAI), a framework designed to facilitate real-time online safety. The system employs a sophisticated ingestion pipeline utilizing Pytesseract for OCR-based text extraction and the Demoji library for graphical icon normalization. A fine-tuned RoBERTa transformer model serves as the core analytical engine, leveraging bidirectional self-attention mechanisms to categorize content across six granular demographic classes with a peak detection accuracy of 96.44%. Beyond mere detection, the framework introduces a Generative AI-driven neutralization module powered by the Gemini-1.5-Flash model. Through targeted prompt engineering, the module identifies toxic linguistic spans and reformulates them into polite, constructive alternatives, successfully reducing toxicity scores from 0.94 to a negligible 0.03 while meticulously preserving the user's original communicative intent. Integrated via a Flask-React architecture with a total end-to-end latency of 450 milliseconds, the proposed system offers a scalable, transparent, and proactive solution for fostering inclusive digital environments.

Key Words: Cyberbullying Detection, Generative AI, Multimodal Fusion, RoBERTa, Natural Language Processing, Semantic Neutralization, Social Media Moderation, Context-Aware AI, Real-Time Monitoring.

1. INTRODUCTION

The rapid proliferation of social media platforms has revolutionized global communication, enabling users to interact through diverse modalities including text, captions, emojis, and high-resolution imagery. While this digital connectivity fosters social engagement, it has simultaneously precipitated a significant rise in cyberbullying and online harassment, frequently targeting women and vulnerable demographics. Modern cyberbullying is no longer confined

to direct offensive text; it manifests through complex channels such as sarcasm, symbolic expressions, and abusive text embedded within memes or screenshots.

Conventional moderation systems and rule-based detection approaches primarily rely on keyword filtering and static text analysis, which are increasingly insufficient to address the contextual nuances of modern digital discourse. These traditional methods often lack the integration required to synthesize text and visual data, resulting in the incomplete detection of harmful interactions. Furthermore, many existing systems are purely reactive, either blocking content entirely or utilizing passive masking techniques that disrupt the conversational flow and lack transparency for the end-user.

Recent advancements in Natural Language Processing (NLP) and transformer-based architectures have enabled more sophisticated sentiment and toxicity analysis. Models such as Bidirectional Encoder Representations from Transformers (BERT) and Robustly Optimized BERT approach (RoBERTa) have demonstrated strong results in capturing bidirectional semantic relationships. Additionally, Generative AI models provide the capability to move beyond mere detection, offering the potential to transform hostile language into constructive communication through automated semantic neutralization. However, many current implementations function as separate tasks, lacking a unified real-time pipeline that combines multimodal extraction, context-aware classification, and active generative correction.

To overcome these limitations, this paper proposes a Multimodal Cyberbullying Detection and Generative Semantic Neutralization System (MCDNS-SMGAI). The framework ingests multimodal inputs—including captions, replies, and text-in-images—utilizing Pytesseract for OCR extraction and the Demoji library for normalization. A fine-tuned RoBERTa classification network analyzes these unified sequences to distinguish between genuine cyberbullying, implicit toxicity, and harmless sarcasm. Flagged content is subsequently processed by a generative language layer that identifies toxic spans and replaces them with polite alternatives using the Gemini-1.5-Flash model. By integrating a multi-model detection pipeline with a generative intervention module, the proposed system offers a proactive and inclusive solution for fostering safer digital environments.

2. LITERATURE REVIEW

Abdullah et al. [1] utilized several machine learning (ML) models, including Support Vector Machines (SVM), Random Forest (RF), and Naïve Bayes, to detect cyberbullying through Natural Language Processing (NLP) cleaning and Term Frequency-Inverse Document Frequency (TF-IDF) feature extraction. While their integration of Bidirectional Encoder Representations from Transformers (BERT) improved contextual accuracy, the system remains text-centric and does not incorporate multimodal support or generative neutralization.

Prama et al. [12] developed an AI-enabled framework using Long Short-Term Memory (LSTM) networks to classify cyberbullying severity into mild and severe levels by combining social media text with user-specific attributes. Despite achieving strong results in severity detection, the reliance on user-specific data raises significant privacy concerns, and the architecture lacks a real-time implementation for immediate content mitigation.

Rahman et al. [13] introduced the CAMFusion framework, which integrates video and textual cues to detect sarcasm and humor through context-aware multimodal fusion. Although this approach successfully extracts visual and semantic representations using deep learning, it is not specifically optimized for cyberbullying detection and involves high computational complexity that limits real-time social media deployment.

Zhang et al. [24] proposed a cross-platform multimodal transfer learning framework that analyzes both text and images across different social platforms to improve detection generalization. While the framework improves accuracy through shared knowledge, it lacks a mechanism for sarcasm detection and does not provide an automated Generative AI (GenAI) module for content detoxification.

Dale et al. [5] explored text detoxification using large pre-trained neural models to rewrite toxic sentences into polite forms through text style transfer. Their research demonstrates that semantic meaning can be preserved while reducing harmful expressions; however, the models operate purely on textual data and are not integrated with a prior cyberbullying detection stage to automate the intervention process.

Yu et al. [23] presented a two-stage detoxification framework utilizing Large Language Model (LLM) fine-tuning to improve model generalization and semantic preservation. While their approach ensures that transformed outputs retain intended meanings, it remains a unimodal, text-only solution that lacks the multimodal support and real-time social media integration necessary for dynamic digital environments.

3. INFERENCE FROM LITERATURE SURVEY

The comprehensive review of existing literature reveals several critical limitations in current cyberbullying detection and content moderation strategies. Existing approaches primarily address isolated components of the problem, such as contextual classification using transformers, imbalanced dataset handling, or text-based style transfer, rather than providing an integrated and unified mitigation framework [1], [5], [12]. Most current detection mechanisms are strictly reactive and text-centric, leading to a failure in capturing multimodal harassment—such as abusive text embedded in images or emojis—resulting in incomplete protection within dynamic social media environments [15], [25].

Machine learning and transformer-based architectures, including BERT and its derivatives, have significantly enhanced semantic understanding and intent classification [7], [14]. However, standalone detection models lack a proactive intervention layer, often defaulting to binary masking or content deletion which disrupts user interaction and communicative flow [11], [16]. Furthermore, no existing approach successfully bridges the gap between high-fidelity multimodal extraction and real-time generative semantic neutralization, where toxic content is not just identified but constructively reformulated in situ [18], [23].

Current activity monitoring and automated moderation methods often operate independently, resulting in a fragmented user experience where the context of a conversation is frequently lost during the interval between detection and manual intervention. Additionally, the lack of sarcasm-aware gating and visual semiotic analysis in traditional models reduces detection accuracy in social media dialects, which are heavily reliant on irony and graphical icons [2], [13], [24].

Based on these identified gaps, this paper proposes the MCDNS-SMGAI Framework—a Multimodal Cyberbullying Detection and Generative Semantic Neutralization System. The proposed system addresses all identified limitations by integrating an OCR-based multimodal ingestion pipeline, a context-aware RoBERTa classification engine, and an active generative intervention module powered by the Gemini-1.5-Flash model. By combining real-time detection with semantic detoxification into a single unified platform, the system shifts the moderation paradigm from passive censorship to inclusive and safe digital communication.

4. PROPOSED SYSTEM

The proposed system introduces the MCDNS-SMGAI Framework, a Multimodal Cyberbullying Detection and Generative Semantic Neutralization System designed specifically for modern social media environments. The framework addresses the escalating psychological risks caused by the proliferation of complex, multi-layered digital harassment. Unlike traditional unimodal and reactive moderation tools, such as the basic machine learning classifiers proposed by Abdullah et al. [1] or the text-only LSTM severity models discussed by Prama et al. [12], the proposed system integrates real-time multimodal data ingestion, context-aware behavioural analysis, and active generative detoxification into a unified platform.

The system continuously processes diverse user-generated inputs, including direct text, captions, graphical emojis, and text-embedded images such as memes or screenshots. These multimodal elements are captured and standardized through an ingestion pipeline utilizing Pytesseract for Optical Character Recognition (OCR) and the Demoji library for visual semiotic decoding. The aggregated data is structured into unified sequence tensors and fed into a dual-gate Context-Aware Detection Engine. This core engine combines a fine-tuned RoBERTa (Robustly Optimized BERT approach) transformer—which generates 768-dimensional contextual embeddings using bidirectional self-attention—with a VADER-based sentiment discordance calculator and a heuristic emoji-mapping dictionary to detect implicit toxicity and masked sarcasm.

The extracted contextual features are processed through a fully connected dense layer and a Softmax classification function, which computes a probability distribution across six granular categories: Age, Ethnicity, Gender, Religion, Other Bullying, and Not Bullying. To minimize false positives, the system employs a rigid probability gating and thresholding mechanism. Sequences with a toxicity probability exceeding 0.8 or a negative sentiment compound score below -0.5 are flagged as high-risk. Conversely, content falling below this critical threshold is classified as standard communication and allowed to pass freely without intervention.

Upon the detection of suspicious or abusive content, the system actively bypasses traditional binary blocking or rudimentary word-masking techniques (e.g., replacing words with asterisks), resolving a major limitation found in existing NLP moderation systems. Instead, it automatically enforces a Generative Semantic Neutralization protocol powered by a Large Language Model (Gemini-1.5-Flash). The generative intervention layer isolates the specific harmful linguistic spans and dynamically substitutes them with polite, constructive good words, meticulously preserving the user's original communicative intent. To ensure transparency and real-time monitoring, the integrated Flask-

React dashboard provides immediate feedback, displaying the classification label, confidence score, and the safely neutralized text. By combining advanced multimodal feature extraction, bidirectional contextual classification, and intent-preserving generative reformulation, the proposed MCDNS-SMGAI framework significantly reduces the psychological impact of digital harassment and fosters a safe, inclusive social media ecosystem.

5. SYSTEM ARCHITECTURE

The MCDNS-SMGAI framework is organized into three functional modules that operate sequentially and collaboratively to deliver end-to-end multimodal data ingestion, context-aware threat detection, and generative semantic neutralization.

5.1 Module I: Multimodal Data Collection and Preprocessing

This module serves as the foundational entry point of the framework, responsible for ingesting multimodal content, establishing a structured data format, and initializing the normalized textual baseline for downstream contextual analysis. The module processes incoming image-based media utilizing Pytesseract Optical Character Recognition to extract embedded text and applies the Demoji library to translate graphical emojis into descriptive equivalents. Extracted text and raw strings undergo rigorous heuristic refinement, including regex-based entity stripping, case normalization, and morphological lemmatization to reduce words to their fundamental root forms. The module accepts raw social media posts, direct text submissions, embedded emojis, and text-embedded image files as input and produces unified, sanitized, lemmatized textual sequences, and aggregated feature vectors for baseline contextual evaluation as output.

5.2 Module II: Cyberbullying and Sarcasm Detection

This module forms the analytical core of the framework. It continuously analyzes the unified text sequences generated from the ingestion layer, extracts contextual embeddings, and passes them through a multi-layered classification architecture to compute precise toxicity probabilities. The standardized textual data, including translated emojis and OCR-extracted strings, are segmented and converted into high-dimensional numerical feature vectors using a RoBERTa-specific sub-word tokenizer and attention masks. These structured tensors are simultaneously processed by a fine-tuned RoBERTa bidirectional transformer, a VADER sentiment discordance engine, and a heuristic emoji-mapping dictionary to capture deep semantic dependencies and masked sarcasm. The extracted contextual features are evaluated by a fully connected dense layer and a Softmax classifier to produce a probability distribution across six

distinct demographic classes: Age, Ethnicity, Gender, Religion, Other Bullying, and Not Bullying. The predicted intent is then evaluated against a rigid confidence gate, triggering active moderation if the toxicity probability exceeds a 0.8 threshold or the sentiment compound score falls below -0.5, immediately initiating the generative neutralization sequence and real-time dashboard alerts.

5.3 Module III: Generative Semantic Neutralization

This module enforces automated content protection controls based on the classification probability received from Module II and provides users with constructive, polite text alternatives for every detected abusive event. Upon receiving a high-risk toxicity score exceeding the 0.8 threshold, the generative intervention engine automatically neutralizes the relevant harmful linguistic spans in real time, substituting offensive words with socially acceptable language using the Gemini-1.5-Flash model. For all flagged events, the integrated Flask-React architecture delivers an instantaneous push update to the user's social media frontend. The generative AI evaluates the underlying contextual meaning of the flagged sentence, producing a refined semantic structure that effectively replaces toxicity while strictly preserving the user's original communicative intent. Instead of presenting only a generic blocked content label, the dashboard seamlessly displays the toxicity class, the probability confidence score, and the safely neutralized plain-language alternative. Administrators and users can use this information to review moderation actions, understand contextual flagging, and maintain uninterrupted, inclusive digital communication across the platform.

6. METHODOLOGY

The MCDNS-SMGAI framework follows a structured methodology that begins with the continuous collection and standardization of multimodal user-generated content across social media platforms. Diverse behavioural inputs including direct text comments, graphical emojis, and text-embedded media such as memes and screenshots are captured per user interaction. These inputs are preprocessed utilizing Pytesseract for OCR-based text extraction and the Demoji library for graphical icon normalization. The aggregated strings undergo rigorous morphological refinement—including regex-based entity stripping and lemmatization—and are subsequently converted into high-dimensional numerical feature vectors using transformer-specific sub-word tokenization and dynamic attention masks.

These unified feature tensors are simultaneously processed by a multi-layered analytical engine to detect nuanced abuse. The core component, a fine-tuned RoBERTa transformer, leverages bidirectional self-attention mechanisms to model complex semantic dependencies and contextual relationships within the text. Concurrently, a VADER-based

sentiment discordance calculator evaluates the emotional polarity of the sequence to detect implicit hostility. A heuristic emoji-mapping module further analyzes the translated graphical icons to pinpoint disguised malicious intent and masked sarcasm that standard global NLP models typically overlook.

The extracted contextual features from these components are aggregated and passed through a fully connected dense layer paired with a Softmax activation function to produce a normalized probability distribution. The content is classified across six specific categories: Age, Ethnicity, Gender, Religion, Other Bullying, and Not Bullying. The computed risk is evaluated against a rigid confidence gating mechanism to determine the intervention tier. A toxicity probability between 0.0 and 0.79, coupled with a neutral or positive sentiment score, is classified as safe and requires no action. Conversely, a toxicity probability of 0.8 or higher, or a negative sentiment compound score below -0.5, is classified as a high-risk violation, triggering full incident logging and immediate automated neutralization enforcement.

Upon high-risk classification, the generative semantic mitigation layer is automatically activated. The framework utilizes the Gemini-1.5-Flash Large Language Model to calculate the contextual weight of the harmful linguistic spans contributing to the high toxicity score. Instead of executing a binary deletion, the generative model actively reformulates the sentence, replacing explicit and offensive terminology with polite, constructive alternatives while strictly preserving the user's original communicative intent. These final outputs are presented instantaneously on the user dashboard via a Flask-React architecture, displaying the specific toxicity class, the probability confidence score, and the safely neutralized text, ensuring transparent, inclusive, and uninterrupted digital communication.

7. CONCLUSIONS

The proposed Multimodal Cyberbullying Detection and Generative Semantic Neutralization System (MCDNS-SMGAI) addresses the escalating psychological risks posed by complex, multi-layered harassment in dynamic social media environments. By integrating a robust multimodal ingestion pipeline with a fine-tuned RoBERTa transformer, VADER sentiment discordance, and an automated generative intervention layer, the system achieves highly accurate detection of cyberbullying across diverse forms of digital communication, including direct text, emojis, and embedded images. The rigid confidence gating and classification mechanism ensures precise, context-aware responses, minimizing false positives and reducing reliance on manual moderation. Real-time semantic neutralization, powered by the Gemini-1.5-Flash model, seamlessly reformulates toxic content into constructive language immediately upon detection, mitigating emotional harm before the end-user is exposed. The integration of this generative capability

transforms traditional, passive censorship into an inclusive, transparent moderation process, improving the digital experience while maintaining an uninterrupted communicative flow.

The framework is highly applicable to diverse digital ecosystems, including mainstream social media platforms, multiplayer online gaming communities, educational discussion forums, and enterprise communication networks requiring continuous content moderation. It effectively supports the protection of vulnerable demographics engaging in online interactions by providing a proactive, automated safety net. The system is highly suitable for social technology conglomerates, educational institutions, digital mental health platforms, and any organization prioritizing safe, inclusive, and constructive digital environments.

Future enhancements may include the integration of advanced cross-lingual models to support real-time cyberbullying detection and neutralization in resource-constrained and regional languages, adaptive generative prompting dynamically tailored to the specific age group or psychological profile of the user, cross-platform API deployment for seamless integration into diverse social media architectures, and the extension of the multimodal pipeline to analyze live audio and video streams. Overall, the framework demonstrates the effectiveness of combining sophisticated multimodal feature extraction, bidirectional transformer analysis, and generative semantic enforcement to achieve a proactive, transparent, and highly scalable solution for digital harassment mitigation in modern social networks.

ACKNOWLEDGEMENT

The authors would like to express sincere gratitude to Dr. G. Santhi, Professor, Department of Information Technology, Puducherry Technological University, for her valuable guidance and continuous support throughout this research work. The authors also thank the Department of Information Technology, Puducherry Technological University, for providing the necessary resources and facilities to carry out this work.

REFERENCES

- [1] M. Abdullah, I. L. Latif, N. Hafeez, F. Ullah, G. Sidorov, E. F. Riverón, and A. Gelbukh, "Cyberbullying Detection on Social Media Using Machine Learning Techniques," *Computación y Sistemas*, vol. 29, no. 3, 2025.
- [2] M. Alharbi, F. Alqahtani, and S. Alsubaie, "Sarcasm detection in online social networks using deep contextual embeddings," *Applied Soft Computing*, vol. 148, p. 110891, 2024.
- [3] F. N. Al-Wesabi, M. Obayya, J. Alsamri, R. Alabdan, N. O. Aljehane, S. Alazwari, F. F. Alruwaili, M. A. Hamza, and A. Swathi, "Automatic Recognition of Cyberbullying in the Web of Things and Social Media Using Deep Learning Framework," *IEEE Transactions on Big Data*, vol. 11, no. 1, pp. 259–270, 2025.
- [4] T. Davidson, D. Warmesley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proc. Int. AAAI Conf. Web Soc. Media*, vol. 11, no. 1, pp. 512–515, 2017.
- [5] D. Dale, L. Li, and L. Mohit, "Text Detoxification using Large Pre-trained Neural Models," *Transactions of the Association for Computational Linguistics*, vol. 9, 2025.
- [6] A. Derbala Yacoub, A. Elsayed Aboutabl, and S. O. Slim, "Multilingual sarcasm detection for enhancing sentimental analysis using deep learning," *J. Commun. Softw. Syst.*, vol. 20, no. 4, pp. 278–280, 2024.
- [7] J. Devlin, M.-W. 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] M. Ibrahim, A. Hassan, and R. Khalifa, "FusionBullyNet: A robust English-Arabic cyberbullying detection framework using dual-encoder transformer architecture," *Mathematics*, vol. 13, p. 172, 2025.
- [9] V. S. Kumar, R. Prakash, and N. Menon, "Real-time cyberbullying detection across multimodal platforms using XLNet," in *Proc. ICAECA*, p. 412, 2025.
- [10] Y. Lu, Y. Zhou, J. Li, Y. Zhang, W. Wang, X. Li, M. Zhang, F. Liu, J. Yu, and M. Zhang, "Adaptive Detoxification: Safeguarding General Capabilities of LLMs through Toxicity-Aware Knowledge Editing," *Findings of ACL*, 2025.
- [11] R. Mishra, S. Patra, and A. K. Tripathy, "A comprehensive survey on cyberbullying detection using natural language processing and deep learning techniques," *IEEE Access*, vol. 12, pp. 11245–11268, 2024.
- [12] T. T. Prama, J. F. Amrin, M. M. Anwar, and I. H. Sarker, "AI-Enabled User-Specific Cyberbullying Severity Detection with Explainability," *IEEE Access*, vol. 13, pp. 12543–12558, 2025.
- [13] M. Rahman, P. K. Dhar, M. A. M. Provath, K. Deb, and T. Shimamura, "Context-Aware Multi-Modal Fusion Framework for Detecting Sarcasm and Humor Integrating Video and Textual Cues," *IEEE Access*, vol. 13, 2025.
- [14] S. K. Roy, P. Dutta, and M. Banerjee, "Text-based cyberbullying detection on social media using transformer-based models," *Expert Systems with Applications*, vol. 235, p. 121012, 2024.
- [15] F. R. Sayed, E. H. Elnashar, and F. A. Omara, "Cyberbullying detection in social media using natural language processing," *Scientific African*, vol. 28, p. e02713, 2025.
- [16] A. Sharma, N. Gupta, and S. Verma, "Cyberbullying detection using NLP techniques," *Int. J. Adv. Res. Comput. Sci.*, vol. 15, no. 2, pp. 45–50, 2023–2024.
- [17] A. Sharma, N. Jain, and V. Kumar, "Detection of abusive and hateful content against women on social media platforms," *J. Intell. Inf. Syst.*, vol. 62, no. 1, pp. 89–110, 2024.
- [18] K. Singh, R. Mehta, and P. Bansal, "Automatic neutralization of toxic language using NLP-based text transformation," *Inf. Manag. Data Insights*, vol. 4, no. 1, p. 100204, 2025.

- [19] X. Song, Y. Liu, and H. Chen, "Assessing the human likeness of AI-generated counterspeech," in Proc. COLING, p. 239, 2025.
- [20] I. Tabassum, S. Rahman, and M. Hossain, "Multilingual multimodal cyberbullying detection through adaptive and hierarchical fusion," IEEE Access, p. 142, 2026.
- [21] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? Predictive features for hate speech detection," in Proc. NAACL Human Lang. Technol., pp. 88–93, 2016.
- [22] Y. Wen, Q. Zhang, and L. Zhao, "DashFusion: Dual-stream alignment with hierarchical bottleneck fusion for multimodal sentiment analysis," IEEE Trans. Neural Netw. Learn. Syst., p. 89, 2025.
- [23] J. Yu, Y. Zhao, J. Zhu, W. Shao, B. Pang, Z. Zhang, and X. Li, "Text Detoxification: Data Efficiency, Semantic Preservation and Model Generalization," in Proc. EMNLP, pp. 32172–32186, 2025.
- [24] W. Zhang, C. Dong, A. Yao, A. Nazari, and A. Gaddam, "Cross-Platform Multi-Modal Transfer Learning Framework for Cyberbullying Detection," Electronics, vol. 15, 2026.
- [25] S. S.-Us-Sakib, M. R. Rahman, M. S. A. Forhad, and M. A. Aziz, "Cyberbullying Detection of Resource-Constrained Language from Social Media Using Transformer-Based Approach," Natural Language Processing Journal, vol. 9, p. 100104, 2024.

BIOGRAPHIES



Dr. G. Santhi holds B.E., M.E., and Ph.D. degrees in CSE, with specialization in CN, IS, Wireless Networks, and CA.



Preethiga P is a B.Tech (IT) student at PTU, currently focused on the fields of AI&ML. She has proficiency in Python, C, C++, Java.



Prema J is a B.Tech (IT) student at PTU, currently focused on the fields of Full Stack Web Development AI&ML. She has proficiency in Python, Java, MySQL



Sudharshan M is a B.Tech (IT) student at PTU, with a keen interest in data-driven system design. He is proficient in programming languages such as Python