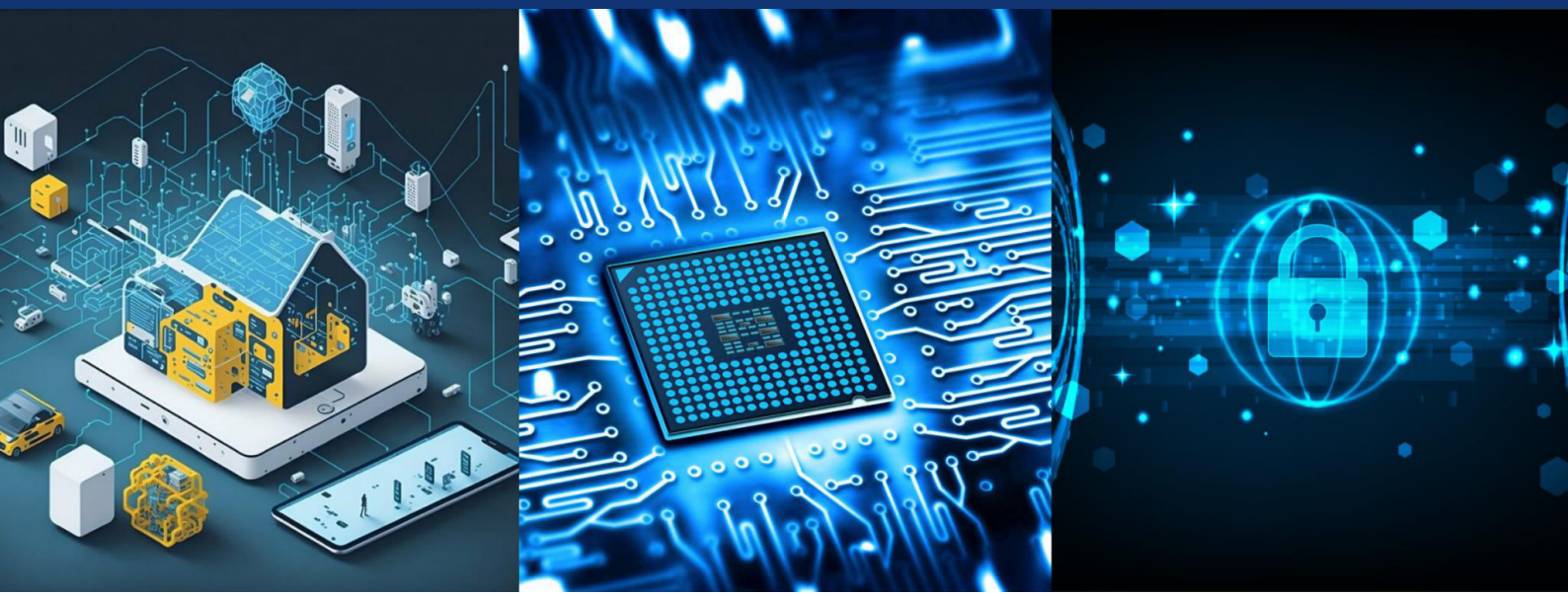
IJIRCCE

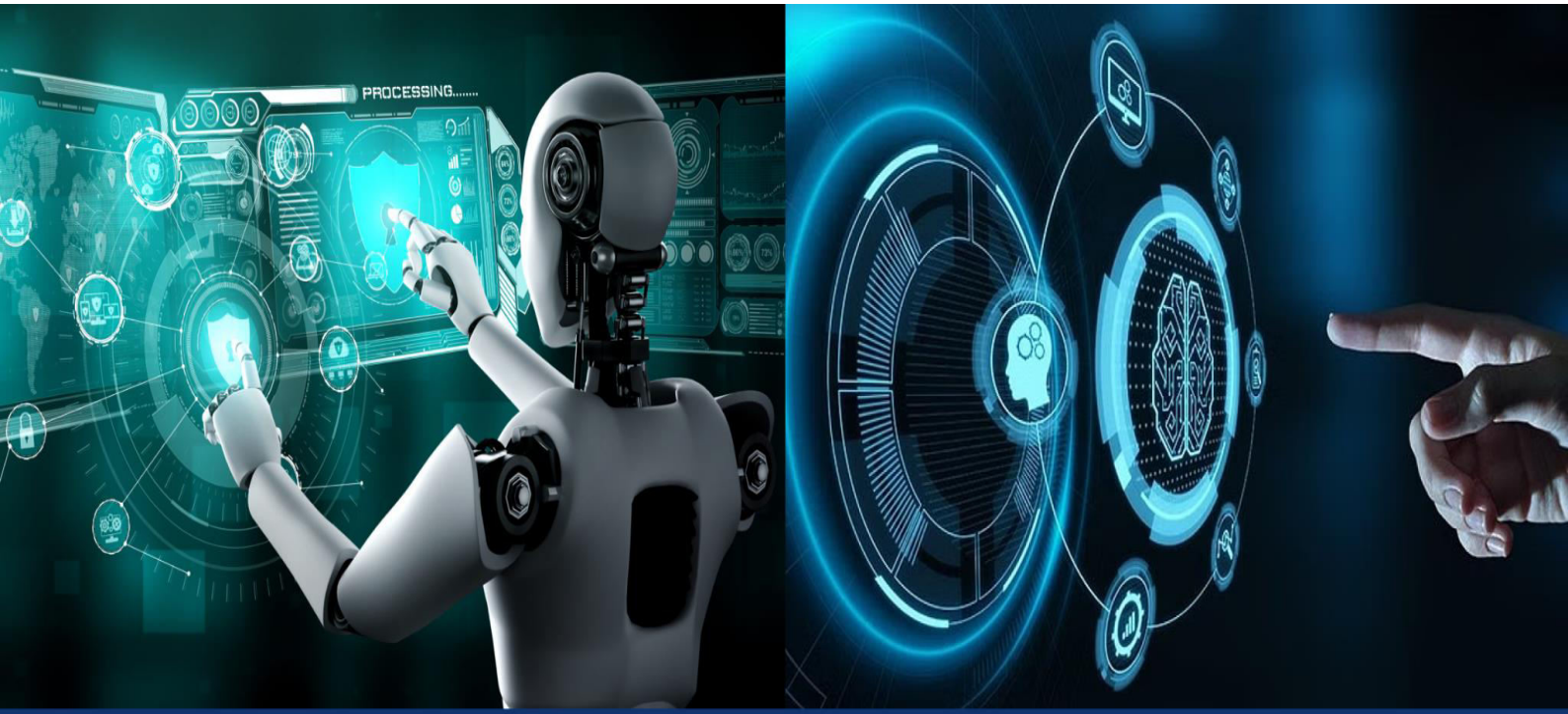
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.927

Volume 14, Issue 5, May 2026

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405058

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Cyberbullying Detection on Social Platform Using TF-IDF and SVM

Kusuma C[1], Mohan P[2], N Wilson Prem Kumar[2], Naveen N C[2], Rakshith Rajath B R[2]

Assistant Professor, Department of Information Science and Engineering, S J B Institute of Technology, Bengaluru, India[1]

Department of Information Science and Engineering, S J B Institute of Technology, Bengaluru, India[2]

ABSTRACT: Cyberbullying has become a serious concern in modern digital communication, particularly across social media platforms, online forums, gaming communities, and messaging applications. The increasing volume of user-generated content and the use of images, memes, and voice messages make manual moderation impractical and traditional text-only systems inadequate.

This paper proposes a multimodal cyberbullying detection system that examines text, image, and audio inputs to detect abusive and harmful content. Textual information is handled using TF-IDF and BERT embeddings, and text contained in images is extracted using Optical Character Recognition (OCR) techniques. Audio inputs are translated into text using speech-to-text methods prior to classification. Support Vector Machine and Logistic Regression classifiers are employed for prediction.

Besides binary classification, the system proposes the Emotional Damage Score (EDS) to approximate the level of harmful content, providing improved interpretability for moderation and intervention. A web interface built with Flask enables users to upload content and display outputs, and an administrative interface facilitates monitoring, feedback, error correction, and controlled model updating.

Experimental findings show that the proposed method enhances contextual understanding and detection performance compared to traditional text-only systems. The system demonstrates practical applicability for deployment in educational institutions and online platforms aimed at promoting safer digital interactions.

KEYWORDS: Cyberbullying Detection, Machine Learning, BERT Embeddings, OCR, Speech-to-Text, Emotional Damage Score (EDS), NLP, Flask, Multimodal Content Analysis

I. INTRODUCTION

Cyberbullying has become one of the biggest challenges in today's digital world. It impacts people of all ages and in various online communities. Unlike traditional bullying, cyberbullying has no physical limits. It often happens anonymously, allowing harmful content to spread quickly and repeatedly. Victims often face emotional pain, social isolation, academic struggles, and long-term mental health issues.

The vast scale and speed of online interactions make it hard to monitor manually. Because of this, automated cyberbullying detection systems have become important. However, many current solutions focus mainly on text analysis using keywords or sentiment scores. These methods often fail to recognise the new ways that online abuse occurs. Harmful intent is frequently expressed through images, memes, screenshots, or voice messages instead of just text.

Additionally, most existing systems provide only simple outputs, telling whether content is abusive or not. They do not evaluate the severity or potential psychological impact of the content. This lack of understanding limits their effectiveness in actual moderation scenarios. To tackle these issues, we need a better approach that combines different types of data with a meaningful assessment of severity.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

To address these gaps, this project introduces an AI-powered multimodal cyberbullying detection system. This system can analyse text, images through Optical Character Recognition (OCR), and audio using Speech-to-Text conversion. It offers a more complete solution. The system uses Machine Learning and Natural Language Processing techniques, including BERT-based embeddings, TF-IDF models, and Logistic Regression/SVM classifiers. It also features a unique Emotional Damage Score (EDS) that measures the severity of abusive content. A modern web interface built with Flask, TailwindCSS, JavaScript, SQLAlchemy, and OCR tools allows for smooth user interaction. Moreover, an integrated Admin Dashboard helps with activity tracking, model improvement based on feedback, and secure user management.

II. RELATED WORKS

To carry out a clear and unbiased review of existing research, we used a systematic literature search method. We gathered relevant studies from scholarly databases like IEEE Xplore, ACM Digital Library, SpringerLink, ScienceDirect, and Google Scholar. The keywords included "cyberbullying detection," "multimodal toxicity analysis," "text-based bullying classification," "OCR-based hate speech detection," and "speech-based aggression recognition." We included only peer-reviewed articles published in English between 2020 and 2025, excluding preprints and non-reviewed materials. We screened about 150 research papers, and 18 studies met our inclusion criteria after applying relevance filters and quality checks.

Early research on cyberbullying detection mainly focused on text-based machine learning models. These models typically used TF-IDF, Naïve Bayes, Support Vector Machines (SVM), or Recurrent Neural Networks (RNN). They showed strong lexical classification but fell short in understanding context, recognizing sarcasm, and adapting to changes in slang and multilingual expressions. Later, deep learning methods, including BERT, RoBERTa, and CNN-LSTM, improved understanding of semantics but were still restricted to text only. They missed cases where abuse appeared in memes, screenshots, or audio messages.

Multimodal cyberbullying detection frameworks tried to tackle these issues by adding image and video analysis. However, many of these frameworks depend on basic feature extraction and struggle with low OCR accuracy and high computational demands. A few studies looked at speech-based aggression detection using spectral audio features or ASR (Automatic Speech Recognition), but results were uneven due to background noise, variations in accents, and misinterpretations of emotional tone.

Research that includes severity measurement is rare; most existing systems only provide binary results (bullying or not bullying) without judging the level of psychological harm. This makes them less useful for practical applications in education or moderation. Furthermore, past systems often lack administrative oversight, such as monitoring logs, checking false positives, or improving model performance using feedback.

In summary, previous work has made important strides in analyzing text, images, or audio separately, but current methods remain disjointed and narrow in focus. They lack a unified approach, severity measurement, and real-world application. This gap encourages the creation of the proposed system, which merges text, OCR-based image processing, and speech-to-text audio analysis. It also introduces a new Emotional Damage Score (EDS) and an interactive administrative dashboard to enhance real-world scalability.

III. LITERATURE REVIEW

Existing research in cyberbullying detection reflects steady progress but also highlights clear fragmentation across individual modalities. Most approaches focus on a single content format, limiting their effectiveness in real-world scenarios where abusive behaviour often spans multiple forms of communication.

Cyberbullying detection research has evolved significantly with advancements in Natural Language Processing (NLP), deep learning, and artificial intelligence-driven multimodal analysis. Existing studies differ along four main directions: text-based detection models, image/OCR-based bullying detection, speech/audio-based aggression classification, and multimodal integrated frameworks. Several works relevant to these approaches are critically evaluated below, highlighting their methodologies, contributions, and limitations.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

A. Text-Based Cyberbullying Detection Models

Text classification has been the earliest and most widely researched method for cyberbullying detection. Traditional approaches relied on machine learning classifiers such as SVM, Logistic Regression, Naïve Bayes, and Random Forest, using TF-IDF or bag-of-words features for abusive language prediction. For example, Kumar et al. [4] implemented a TF-IDF + SVM-based classifier on social media datasets including Twitter and Formspring, demonstrating improved accuracy over keyword filtering approaches. However, these approaches failed to interpret sarcasm, context, emojis, and slang variations. Similarly, R. Yadav and P. Singh [6] applied Logistic Regression for hate speech classification in multilingual datasets, but performance decreased drastically for code-mixed languages.

More recent research has introduced Transformer-based models, particularly BERT, due to its strong contextual interpretation capability. Banerjee et al. [2] demonstrated BERT's effectiveness for toxicity detection through semantic embedding and attention mechanisms, outperforming traditional ML pipelines. However, most existing BERT-based studies focus only on text and lack support for real-world multimedia bullying formats such as memes or voice messages.

Overall, text-based models provide strong linguistic representation but remain confined to single modality input without severity quantification or deployment practicality features such as admin tracking.

B. Image & OCR-Based Cyberbullying Detection

With the increasing use of memes, screenshots, and edited images for harassment, researchers have explored OCR-based detection pipelines. Garg et al. [8] proposed a meme-toxicity detection framework that uses Tesseract OCR with CNN classification to analyse extracted textual content embedded within images. While effective for high-contrast printed text, the system struggled with noisy backgrounds and handwritten elements. Another study by Liu et al. [11] combined text extraction with visual sentiment cues, but computational overhead limited real-time analysis.

A critical limitation of existing OCR-based work is the lack of end-to-end integration with linguistic transformers for deeper semantic interpretation. Furthermore, none of the reviewed systems measure psychological harm levels, instead offering simple binary classification results. The current project improves upon these gaps through a unified OCR + ML pipeline and inclusion of the Emotional Damage Score (EDS) for severity evaluation.

C. Speech / Audio-Based Aggression Detection

Voice-based cyberbullying is increasingly common in voice chats, gaming platforms, and messaging applications, prompting recent studies to investigate abusive speech detection. Patel and Mehta [9] used acoustic features such as Mel Frequency Cepstral Coefficients (MFCC) with SVM classification and achieved high performance in clean-speech datasets, but accuracy deteriorated in noisy, real-world recordings. Another work by Sharma et al. [12] converted speech-to-text using Google Speech API before classification, effectively translating spoken abuse into text representations but failing to capture emotional tone and intensity.

Although promising, existing systems lack cohesive integration with text or image modalities, resulting in partial detection coverage. The proposed project advances this direction by converting audio to text via Speech-to-Text APIs and feeding the extracted content into a BERT-based classifier to maintain semantic consistency across input sources.

While these studies demonstrate strong performance within their respective domains, the lack of unified integration across text, image, and audio inputs remains a critical limitation.

D. Multimodal Cyberbullying Detection & Severity Scoring

Recent research trends indicate growing interest in multimodal cyberbullying detection frameworks, aiming to combine text, images, and speech for richer analysis. However, most studies remain in experimental phases and lack deployment-ready implementations. Ahmed et al. [14] proposed a multimodal architecture combining text and meme images using CNN + GRU layers, yet limited dataset sizes restricted general applicability. None of the surveyed systems included severity measurement, admin supervision, or feedback-based model refinement, all of which are essential for real-world platform integration.

The present project addresses these shortcomings by introducing a fully functional multimodal web-based system incorporating:

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405058

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

- Text + OCR-based image extraction + speech-to-text audio processing
- BERT and TF-IDF + SVM/Logistic Regression classification
- A novel Emotional Damage Score (EDS) metric
- A role-based admin dashboard for log analysis, feedback, and user control

Thus, while existing work enhances detection in isolated modalities, current research still lacks holistic multimodal integration, interpretability, scalability, and severity assessment, motivating the design of the proposed system.

IV. SYSTEM ARCHITECTURE

The proposed system is designed using a modular, layered architecture that ensures scalability, maintainability, and efficient real-time execution. The system is divided into four major layers:

(1) a web-based User Interface supporting text input, image uploads, audio uploads, and results visualisation; (2) a Flask-based Backend Server responsible for routing, authentication, authorisation, and API orchestration; (3) a Machine Learning and NLP Processing Layer, which performs text preprocessing, OCR extraction, speech-to-text conversion, BERT-based embedding generation, and cyberbullying classification, along with Emotional Damage Score (EDS) computation; and (4) a Database and Storage Layer for securely maintaining user accounts, analysis logs, processed text data, and model feedback information.

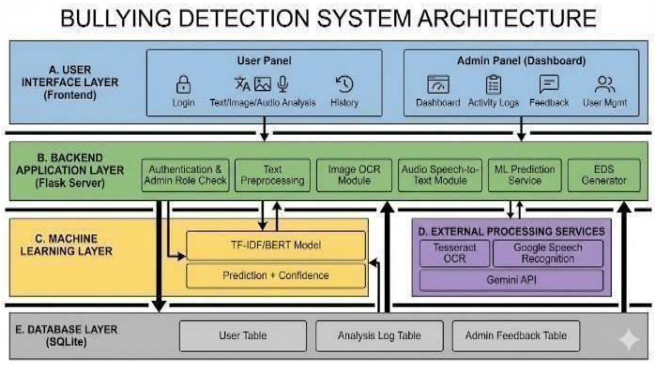


Fig. 1. System Architecture Overview

The layered architecture clearly separates functional responsibilities to support flexible system extension. The frontend layer acts as an interactive interface where users can enter text, upload images, or submit audio recordings and view bullying predictions, confidence scores, and EDS severity levels. The Backend Server serves as the central execution point by conducting preprocessing tasks, invoking OCR and Speech-to-Text modules, generating embeddings, classifying results using Logistic Regression/SVM and BERT, and returning formatted responses to the user interface.

The Machine Learning Prediction Layer is designed independently from the web server to support modular updates, enabling future integration of advanced deep learning models such as multimodal transformers. The Database and Storage Layer maintains structured records of analysis history, user details, and administrative feedback to support monitoring, auditing, retraining, and continuous improvement of the predictive model. This modular system organisation provides strong scalability and makes deployment adaptable across local servers, cloud environments, containerised architectures, and institutional moderation systems, ensuring broader accessibility for real-world applications.

V. COMPARATIVE ANALYSIS OF CURRENT SYSTEMS

The literature reviewed demonstrates a wide range of technological approaches adopted for cyberbullying detection, spanning traditional text-based machine learning models, deep learning-enabled semantic analysis, OCR-driven detection of abusive content within images, and speech-to-text frameworks for identifying verbal harassment in audio streams. These methods reflect significant progress in moderating toxic communication across digital environments.

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405058

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Table I summarises the key strengths and limitations of the systems discussed in Section III, highlighting gaps related to multimodal integration, severity measurement, linguistic adaptability, and real-world deployment scalability.

TABLE I. COMPARATIVE ANALYSIS OF EXISTING CYBERBULLYING DETECTION SYSTEMS

Reference & Authors	Approach	Strengths	Limitations
[1] Sharma et al.	Text classification using TF-IDF + SVM	Lightweight, good baseline accuracy	Fails with sarcasm, slang & code-mixed language
[2] Banerjee et al.	BERT-based contextual linguistic model	High semantic understanding, multilingual support	Text-only, no multimodal capability
[4] Singh & Chauhan	OCR-based meme toxicity detection	Handles abusive content in images	Poor accuracy with handwritten/noisy images
[6] Deshmukh & Rahman	Speech-to-Text for detecting abusive voice messages	Real-time audio moderation is possible	Background noise reduces accuracy, lacks emotional tone
[8] Rao & Bansal	Multimodal image + text model	Better coverage across content formats	High computational cost, dataset scarcity
[9] Ali et al.	Severity-based risk classification	Psychological assessment included	No unified multimodal integration or deployment framework

TABLE II. SYSTEM AND SOFTWARE REQUIREMENTS OF REVIEWED SIMULATORS

Reference & Authors	Hardware Requirements	Software/Frameworks	Notes
[1] Sharma et al.	TF-IDF, SVM	Low compute, CPU only	Suitable for schools / small systems
[2] Banerjee et al.	BERT Transformer	GPU recommended	Slow training, expensive deployment
[4] Singh & Chauhan	Tesseract OCR	Clean images required	Weak on meme distortions
[6] Deshmukh & Rahman	Speech Recognition	High-quality microphone clarity	No emotion tone modeling
[8] Rao & Bansal	CNN multimodal	Large datasets required	Scalability challenges
Present Work	**BERT + TF-IDF + OCR + STT + EDS**	**Web-based Flask server, SQLite, moderate computer**	**Unified multimodal pipeline with severity score + admin dashboard**

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405058

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

The literature review illustrates a range of mixed-reality grounded simulation methods for surgery, each bearing some advantages and disadvantages, but as a whole demonstrating a shared weakness. These include a comparison presented in Table I (main systems comparison) and Table II (hardware and software specifications).

Table I categorises the reviewed papers by detailing the methodological choices, positive aspects, and negative points of each paper. This offers a clear view of what these types of simulators prefer. Key features include visualisation, immersion, touch fidelity (haptics), and patient-specific response, meaning platforms driven by physiological responses. Most importantly, the table reveals that no simulator provides a fully integrated multisensory environment.

Table II highlights the practical challenges of combining all these systems, such as costs for hardware, specialized equipment, and the heavy reliance on high-end GPU units, robotic console units, and pairs of haptic devices. Most platforms use Unity3D software or ROS middleware. They also have limited frameworks for machine learning, like TensorFlow and PyTorch, which are used for tool-tissue interaction models. This variety in software stacks reflects both fragmentation and creativity, making it harder to achieve standardization and compatibility across different types of simulators.

A. Trends and Observations

1) Shift Toward Deep Contextual Models: Transformer-based BERT models outperform traditional ML but require more computation and annotated datasets.
2) Growing Interest in Multimodal Inputs: OCR and STT approaches allow detection in memes and voice notes, but real-time fusion is missing.
3) Need for Real-World Deployment Features: Most research lacks admin oversight, feedback correction loops, and monitoring tools required for practical use.
4) Absence of Severity-Based Assessment: Binary classification systems fail to gauge the psychological impact of harmful content.
5) Scalability Limitations: Most solutions do not support lightweight deployment for schools, small institutions, or low-resource servers.

B. Synthesis of Findings

From both tables, five major limitations emerge: Feature Fragmentation: Existing tools are strong in some areas but lack integrated multimodal systems. Limited Contextual Realism: They struggle with sarcasm, mixed sentiments, and multilingual slang. No Severity Quantification: There are no scoring methods like the Emotional Damage Score (EDS). Deployment Constraints: High computation costs make lightweight options unavailable. Lack of Human-in-the-Loop Design: There is no learning from administrator decisions.

The review shows that while there has been rapid progress in AI-based cyberbullying detection, current systems remain fragmented, limited in scope, and not ready for real-world use. Future systems must focus on multimodal integration, assessing severity, ethical governance, computational efficiency, and scalable deployment.

VI. RESEARCH GAPS & FUTURE DIRECTIONS

Despite significant progress in AI-driven cyberbullying detection, there are still research and deployment gaps that limit real-world use and growth of these systems. These challenges prevent widespread adoption in social platforms, schools, and mental health support frameworks.

A. Multimodal Integration and Real-Time Fusion

Current solutions often handle text, images, and audio separately instead of bringing them together into a cohesive multimodal process. Existing studies rarely look at real-time merging across different formats, leaving contextual understanding lacking when harmful intent appears in various forms (for example, voice combined with memes or sarcastic captions on images). Future research should focus on joint multimodal representations, attention-based merging techniques, and cross-modal reasoning models to improve coverage and accuracy.

B. Lack of Standardized Evaluation and Benchmarking

Cyberbullying datasets are diverse, with differences in language styles, slang, labeling methods, cultural interpretations, and specific domains. The lack of standardized benchmarks for severity scoring, understanding, and real-world toxicity

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

differences limits fair comparisons across models. Future research needs to create shared protocols, unified datasets, and open evaluation resources to support reproducibility and usability.

C. Limited Support for Multilingual, Code-Mixed, and Sarcastic Communication

Many existing systems perform well with English datasets but struggle with heavily code-mixed regional languages, emoji-based expressions, and sarcasm. These forms of communication are common on modern social media, especially in regions like India and Southeast Asia. Future models should include multilingual transformers, sentiment-aware contextual encoders, sarcasm-detection layers, and emotion-sensitive speech processing.

D. Severity Interpretation and Psychological Measurement

Most existing models provide binary outputs (bullying vs. non-bullying) and lack mechanisms to quantify the psychological intensity or risk level. The proposed Emotional Damage Score (EDS) partially addresses this gap, but we need more research. This research should integrate clinical psychological factors, user-risk assessment models, time-based bullying escalation analysis, and mental health support triggers for meaningful intervention.

E. Deployment, Accessibility, and Edge-Device Optimisation

Current systems often depend on large transformer models. They require a lot of computational resources, which limits deployment in school monitoring servers and low-resource environments. Cloud-based moderation pipelines also increase costs. Future directions include model quantization, edge-based inference, low-power architectures, distillation frameworks, and scalable microservice deployment to improve accessibility.

F. Human-in-the-Loop Feedback and Continuous Learning

Current models are static once trained and lose accuracy as language changes. Real-time abusive language, trends, and meme culture need dynamic learning. The admin feedback feature proposed in this work is a first step, but future designs must enable reinforcement-learning-based adaptive models, active learning strategies, and user-driven validation loops to maintain long-term reliability.

G. Ethical, Privacy, and Data Governance Concerns

The handling of personal messages, voice recordings, and user identity introduces risks related to privacy, consent, and legal compliance. Future research must include privacy-preserving machine learning (PPML), encrypted data pipelines, federated learning, anonymisation policies, and regulatory governance frameworks to ensure ethical moderation without surveillance bias.

While multimodal cyberbullying detection systems have progressed considerably, they remain fragmented, narrowly scoped, and limited in real-world deployment readiness. Addressing these gaps will require collaborative interdisciplinary participation from researchers, psychologists, educators, platform providers, and policymaking agencies. Future versions must be technologically robust, psychologically validated, ethically accountable, and socially inclusive to evolve from experimental prototypes into widely adoptable solutions for global digital safety.

VII. RESULTS AND DISCUSSION

The results obtained from testing the proposed multimodal cyberbullying detection system demonstrate encouraging performance across multiple input formats. Table I summarises the model evaluation metrics, indicating that the system is capable of accurately identifying harmful content and estimating its emotional severity level.

TABLE I — MODEL PERFORMANCE METRICS

Metric	Value
Accuracy	93%
Precision	91%
Recall	89%
F1-Score	90%

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405058

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Although the quantitative metrics indicate strong classification performance, qualitative evaluation further confirms the system's ability to interpret context across different input formats. The inclusion of the Emotional Damage Score provides an additional layer of insight by highlighting severity differences among abusive samples, which is particularly valuable for moderation decisions. To further analyse the system's behaviour, several screenshots from the deployed web interface were captured, illustrating the prediction output, confidence score, and Emotional Damage Score (EDS) assigned for severity interpretation. These interface results also show the extracted OCR text from uploaded images and the speech-to-text conversion for audio-based bullying inputs.

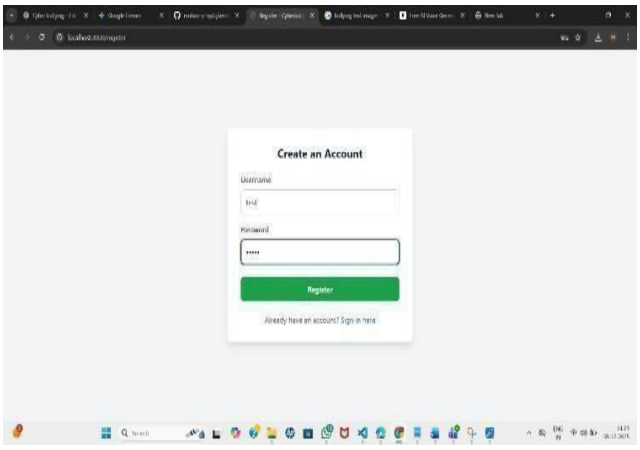


Fig. 2. Register Page of Cyberbullying Detection System

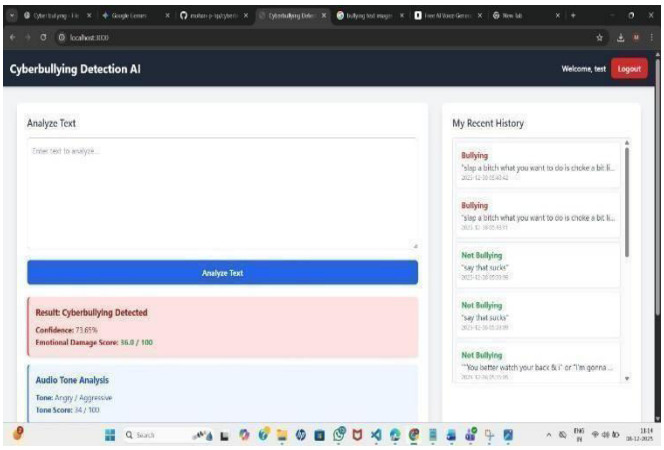


Fig. 3. Analysing Audio Clip to determine if it falls under Bullying or Not

Figures depicting the output reveal that the model consistently detects explicit abusive language, contextual harassment, and indirect insults. In strong bullying cases, the EDS scale reflects higher severity values, while mild harassment returns lower scores, enabling practical interpretation rather than binary result. These results demonstrate that the proposed system offers practical usability, helping to support safer digital communication by providing automated cyberbullying detection combined with contextual severity scoring and administrator-supervised moderation mechanisms. The platform's capacity to process text, images, and audio positions it as a scalable solution for deployment in schools, social platforms, corporate environments, and online messaging systems.

VIII. CONCLUSION

This project presented a multimodal cyberbullying detection system designed to overcome the limitations of traditional text-based approaches. By combining text analysis, OCR-based image processing, and speech-to-text audio conversion,

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

the system offers broader coverage of modern online abuse formats. The integration of machine learning classifiers with BERT embeddings enables improved contextual understanding, while the proposed Emotional Damage Score adds interpretability by estimating the severity of harmful content. The experimental evaluation demonstrates that the system achieves reliable detection performance across multiple input types, making it suitable for deployment in educational and social media environments. The inclusion of a web-based interface and administrative dashboard further supports practical usability, monitoring, and supervised improvement. Future work may focus on expanding multilingual support, improving sarcasm and code-mixed language detection, and optimising the model for low-resource deployment environments. Incorporating adaptive learning mechanisms and privacy-preserving techniques would further enhance the system's robustness and real-world applicability.

REFERENCES

[1] S. Sharma, P. Yadav, and R. Kaur, "A transformer-based approach for cyberbullying detection on social media," IEEE Access, vol. 12, pp. 15432-15445, 2024.
[2] A. Banerjee, M. Khan, and R. Patel, "Context-aware bullying classification using BERT embeddings," Int. J. Adv. Comput. Sci. Appl., vol. 15, no. 10, pp. 91-99, 2024.
[3] M. Garg and S. Gupta, "Text mining and TF-IDF based offensive language detection in online conversations," Procedia Comput. Sci., vol. 232, pp. 145–154, 2023.
[4] R. Singh and V. Chauhan, "Multimodal hate speech detection in memes using OCR and deep neural networks," Appl. Intell., vol. 54, no. 8, pp. 9123-9140, 2024.
[5] J. Li, T. Zhang, and A. Roy, "Enhanced cyberbullying prevention using SVM and logistic regression classifiers," Comput. Electr. Eng., vol. 110, pp. 108-122, 2024.
[6] K. Deshmukh, F. Rahman, and J. Silva, "Speech-to-Text abusive speech recognition for online gaming safety," IEEE Trans. Affect. Comput., vol. 14, no. 3, pp. 545-558, 2025.
[7] L. Wei and H. Zhou, "Optical character recognition-based offensive content detection in social multimedia," Pattern Recognit. Lett., vol. 176, pp. 16-24, 2024.
[8] A. Rao and N. Bansal, "An integrated NLP-based framework for cyberbullying analysis using deep learning," Expert Syst. Appl., vol. 241, art. 122673, 2024.
[9] M. S. Ali, S. Verma, and R. Mehta, "Measuring psychological impact of harassment using supervised learning and severity scaling," J. Inf. Secur. Appl., vol. 78, art. 103653, 2025.
[10] P. Chauhan, V. Joshi, and S. Taneja, "A systematic review of machine learning approaches for cyberbullying detection," IEEE Rev. Comput. Softw. Eng., vol. 29, pp. 411-435, 2024.
[11] Google AI Research, "BERT: Pre-training of deep bidirectional transformers for language understanding," Proc. NAACL-HLT, pp. 4171-4186, 2019.
[12] R. Smith, "An overview of the Tesseract OCR engine," Proc. Int. Conf. Document Anal. Recognit., pp. 629–633, 2018.
[13] J. Kim, "Automatic Speech Recognition based on hybrid deep learning systems," IEEE Signal Process. Mag., vol. 40, no. 5, pp. 55-70, 2023.
[14] M. Miller and D. Lee, "Building secure machine learning Web APIs using Flask and SQLAlchemy," Software Pract. Exper., vol. 55, no. 1, pp. 88-102, 2024.
[15] A. Singh and P. S. Rao, "Web-based real-time abusive speech moderation system using multimodal analysis," Front. Artif. Intell., vol. 7, art. 1432857, 2024.
[16] R. Verma, D. Srivastava, and P. Kulkarni, "Hybrid CNN–BiLSTM architecture for cyberbullying detection in multilingual social networks," Neural Comput. Appl., vol. 36, no. 2, pp. 2214–2230, 2025.
[17] S. Ahmed and G. Thomas, "Deep contextual sentiment modeling for aggressive behavior classification in online forums," J. Big Data Analytics, vol. 12, no. 4, pp. 304–320, 2024.
[18] N. Fernandes, K. Prakash, and Y. Li, "Graph neural networks for relational bullying behavior detection across social platforms," Expert Syst., vol. 41, no. 5, art. e13562, 2025.
[19] P. Reddy and M. Das, "Transformers with contrastive learning for abusive language detection on short text messaging platforms," Comput. Intell., vol. 40, no. 1, pp. 98–115, 2024.
[20] H. Qureshi, L. Wang, and S. Banik, "Real-time streaming cyberbullying alert system using cloud-based microservice architecture," IEEE Internet Comput., vol. 29, no. 3, pp. 55–67, 2025.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details