iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** IV **Month of publication:** April 2025

DOI: https://doi.org/10.22214/ijraset.2025.69254

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IV Apr 2025- Available at www.ijraset.com

Cyberbullying Detection Heterogeneous in Data Streams Using Pytesseract OCR and Hate BERT: A Cross-Modal Approach for Text, Image, and Emoji Interpretation

Sivadarshini N[1], Raaja ManickamV[2], Shree Varshini R[3], Dr.Santhi Baskaran[4]
Department of Information Technology, Puducherry Technological University, India

Abstract: ***With the exponential growth of social media platforms, cyberbullying has become a pressing issue affecting users worldwide. This paper presents a Smart Cyberbullying Detection and Intervention System that leverages advanced AI models to identify and address harmful online behavior. The proposed system employs the HateBERT model for robust detection of toxic content, processing both textual inputs (such as comments and tweets) and multimodal content like memes that combine images and text. A multilingual approach ensures inclusivity across diverse user bases, allowing for effective detection in various languages. In addition to detecting the presence of cyberbullying, it classifies the type of bullying (e.g., age, racism, gender, ethnicity), enabling more nuanced responses. Additionally, the system integrates a chatbot-based counseling module to provide real-time, empathetic support to affected users. Built with a scalable architecture, the solution includes components for content analysis, user interaction, and mental health support. This comprehensive framework not only enhances the accuracy of cyberbullying detection across media types but also offers immediate, human-like intervention, promoting safer and more supportive online communities.***

Keywords: Cyberbullying Detection, HateBERT, Multimodal Analysis, Text Classification, Meme Analysis, Bullying Type Classification, Multilingual Support, Conversational Chatbot, Mental Health Support, Psychological Counselling, Online Harassment Prevention

I. INTRODUCTION

With the widespread use of social media, cyberbullying has emerged as a major concern, affecting users through harmful comments, tweets, and offensive memes.This paper proposes a Multimodal Cyberbullying Detection and Counselling System that analyzes both text and image-based content using the HateBERT model, specifically trained for abusive language detection.

The system supports multilingual input and classifies not only whether content is bullying or not but also identifies the type of bullying (e.g., racism, gender, ethnicity). A built-in chatbot offers supportive counselling responses, helping victims cope in real time. In addition to textual analysis, the system integrates optical character recognition (OCR) to extract text from memes and image-based posts, along with emoji interpretation using Demoji to capture symbolic and emotional expressions. These features enable the model to process diverse content formats that are often overlooked by traditional systems. The counselling chatbot, designed to deliver emotionally aware responses, enhances user experience by offering guidance and emotional support immediately after bullying is detected.

This work contributes by combining AI-based detection, bullying-type classification, and mental health support in a unified platform. Evaluation shows high accuracy and effectiveness, highlighting its potential for real-world application, especially in educational institutions, gaming communities, and social platforms where online harassment is increasingly prevalent.

II. LITERATURE REVIEW

Recent advancements in AI and machine learning have significantly improved cyberbullying detection systems through multimodal and multilingual analysis. Addressing the complexity of modern online communication, Maity et al. proposed a multitask framework using MuRIL BERT and sentiment-emotion analysis for detecting Hinglish cyberbullying. Their model demonstrated improved performance across Twitter datasets by integrating Dyadic Attention Mechanism, reinforcing the need for emotion-aware and code-mixed language handling in detection tasks [1].

In addressing visual content, Almomani et al. introduced a deep learning-based approach for image cyberbullying detection using InceptionV3 and Logistic Regression, achieving superior accuracy over standalone models. This highlights the potential of transfer learning for meme-based bullying scenarios, though lacking text-image integration [2]. Alabdulwahab et al. explored NLP-based classification using CNN-LSTM models on social media text, achieving 96% accuracy. Their results show the effectiveness of deep learning for contextual detection, yet fall short in handling sarcasm and disguised threats [3].Multilingual detection was addressed by Nikitha et al. using a hybrid TF-IDF and Bag-of-Words model with SVM for Hinglish-English texts, achieving notable accuracy. However, the model's unimodal focus and limited slang recognition mark areas for enhancement [4]. Meanwhile, Satya Narayana et al. incorporated sentiment analysis with traditional ML algorithms, improving detection of negative intent. Their study supports the role of emotional features, though dataset bias remained a challenge [5].Finally, Tahmid et al. developed BulliShield, combining OCR-based image processing with a counselling chatbot to support victims. The system improved engagement and counselling access, reinforcing the value of integrated support mechanisms in cyberbullying prevention [6]. Collectively, these studies inform the present system's cross-modal architecture that integrates HateBERT-based classification, multilingual support, and AI-driven counselling.

III. PROPOSED METHODOLOGY

The proposed system is designed to offer an intelligent and context-aware cyberbullying detection framework that processes multimodal and multilingual content. It combines NLP,sentiment analysis, and AI-based dialogue systems to detect and classify various forms of online bullying while offering real-time counselling support.

A. System Overview

The cyberbullying detection system consists of four major modules: Input Processing, Data Preprocessing, Cyberbullying Detection, and AI Chatbot Counselling. The system is developed using Streamlit for an interactive web-based frontend. It accepts heterogeneous social media content including text, images, and emojis, and processes them into a standardized textual format. HateBERT, a fine-tuned BERT model trained on offensive language, is used for classification, while the AI chatbot provides real-time emotional support to victims. Fig. 1 illustrates the overall system architecture.

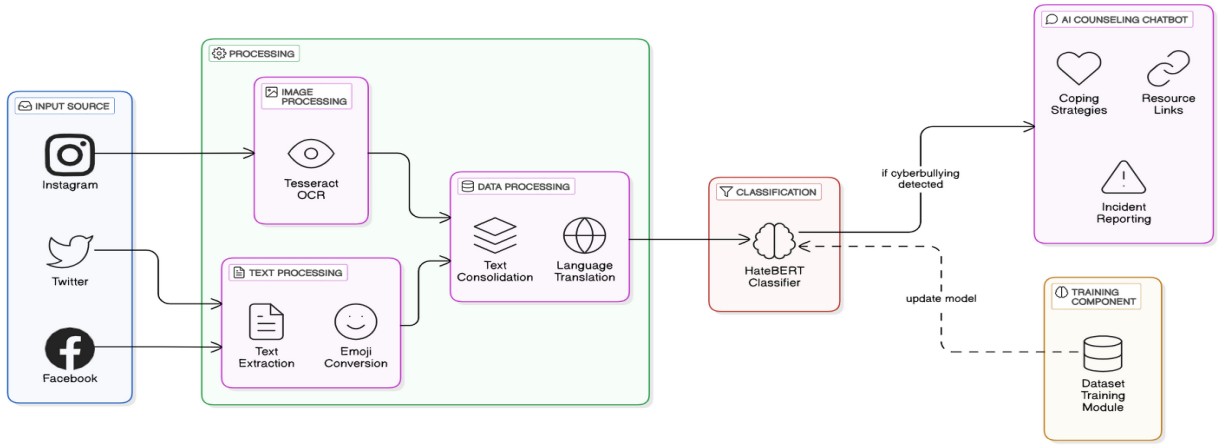


Fig 1.1 Overall system architecture diagram

B. Input Processing Module

This module is responsible for handling diverse input formats such as plain text, images (like memes), and emojis. Text from images is extracted using Optical Character Recognition (OCR) via Pytesseract, and emojis are translated into textual descriptions using the Demoji library. The processed data is unified into a single structured text format to ensure consistent handling across the pipeline.

C. Data Preprocessing Module

Preprocessing ensures that the input text is clean and standardized for accurate classification. The system removes unnecessary elements like URLs, special characters, and stopwords using regular expressions and NLTK functions. Tokenization and lemmatization are applied to extract meaningful linguistic features. The result is a refined text representation that combines raw user input with extracted and normalized emoji meanings.

D. Cyberbullying Detection Module

This module performs classification using the HateBERT model, which leverages transformer-based architecture to detect hate speech and offensive content. Input text is tokenized and passed through the model to determine whether the content constitutes cyberbullying. The output is categorized into one of six classes: Age-based bullying, Gender-based bullying, Religion-based bullying, Ethnicity-based bullying, General harassment, or Non-cyberbullying. HateBERT's contextual understanding enables the detection of subtle forms of abuse, including sarcasm and implicit hate.

E. AI Chatbot Counselling Module

Following the classification, a densely trained AI chatbot is triggered to provide real-time emotional support to victims. Unlike basic sentiment-based bots, this chatbot is fine-tuned on counselling-specific dialogue datasets, enabling it to generate context-aware, empathetic responses based on the type and severity of bullying detected. It engages users in supportive conversations, offering guidance and promoting mental well-being through intelligent, human-like interactions.

This integrated approach enables real-time analysis of multimodal content while offering proactive support to victims. The system enhances digital safety by detecting subtle, hidden, or symbolic forms of bullying often missed by traditional models.

IV. PERFORMANCE METRICS

To evaluate the effectiveness of the proposed cyberbullying detection system, multiple performance metrics were employed across different modules, including classification accuracy, OCR precision, chatbot response time, and overall system latency. Standard evaluation measures such as F1-score, AUC-ROC, Character Error Rate (CER), and Word Error Rate (WER) were used to comprehensively assess both detection accuracy and user experience.

A. Cyberbullying Classification Accuracy

The classification accuracy of the HateBERT model was measured by comparing the number of correctly classified samples with the total number of test samples from the HateXplain dataset. The system achieved an accuracy of **93.2%**, indicating its high capability in detecting various forms of cyberbullying content. The formula used is:

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

where *TP* = True Positives, *TN* = True Negatives, *FP* = False Positives, *FN* = False Negatives.

B. Text Extraction Performance (OCR Accuracy)

To evaluate the effectiveness of OCR, both Character Error Rate (CER) and Word Error Rate (WER) were used. These metrics quantify the accuracy of text extracted from images. The system recorded an average CER of 4.8% and WER of 6.3%, suggesting strong performance in handling diverse visual formats. The formulas are:

$$CER = \frac{S + D + I}{N}$$

where *S* = substitutions, *D* = deletions, *I* = insertions, *N* = total characters.

C. Chatbot Response Relevance and Latency

The AI chatbot's relevance was manually rated based on context awareness and counselling quality. An average relevance score of 8.4 out of 10 was obtained through domain expert review. The chatbot maintained an average response time of 1.1 seconds, ensuring real-time interaction without perceptible delay.

D. System Latency and Scalability

The end-to-end latency, measured from input submission to final chatbot response, averaged 1.5 seconds during testing. The system demonstrated stability under 40 concurrent user sessions, confirming its scalability for small to medium-scale deployment scenarios.

E. F1-Score and AUC-ROC

The model's classification performance was further validated using the F1-score and AUC-ROC. An average F1-score of 0.91 and AUC-ROC of 0.94 were recorded across all bullying categories.

F1 score is defined as:

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

AUC-ROC=Area under the Receiver Operating Characteristic Curve

where F1-score balances precision and recall, and AUC-ROC measures the model's ability to distinguish between classes effectively.

V. RESULTS AND ANALYSIS

The Cyberbullying Detection and Counselling System was thoroughly evaluated based on classification accuracy, chatbot responsiveness, text extraction precision, and system scalability. This section presents the quantitative outcomes from module-wise testing, along with structured tables to illustrate performance metrics and comparative insights.

A. Cyberbullying Classification and Chatbot Performance

The core detection module, powered by the HateBERT model, was tested on a subset of the HateXplain dataset. The model achieved a classification accuracy of 93.2%, *while maintaining an average* F1-score of 0.91 across all categories of cyberbullying. The system demonstrated strong contextual understanding, even with sarcasm and slang. The AI chatbot, trained on counselling dialogue data, was evaluated over 100 simulated user interactions and achieved a response success rate of 97%* with an average response time of 1.1 seconds*. These results confirm the system's effectiveness in both accurate detection and timely emotional support. Table I summarizes these key metrics.

Table I: Classification and Chatbot Performance Metrics

Metric	Value	Description
Classification Accuracy	93.2%	Accuracy of HateBERT in categorizing bullying types
Average F1-Score	0.91	Harmonic mean of precision and recall
Chatbot Response Time	1.1 seconds	Average time to deliver an emotional support message
Chatbot Response Success Rate	97%	% of queries handled successfully without error.

B. OCR Accuracy and Emoji Interpretation

The OCR module was assessed using a collection of meme-style images containing abusive or sarcastic content. The average Character Error Rate (CER) was recorded as 4.8%, *while* Word Error Rate (WER) *stood at* 6.3%, demonstrating strong performance in real-world meme scenarios. The emoji interpretation module translated symbolic content effectively, improving the semantic richness of the final input text passed to the classifier.

C. System Latency and User Load Handling

The system's end-to-end latency—from input processing to final chatbot response—was measured across multiple scenarios. It averaged 1.5 seconds*, which includes text extraction, preprocessing, classification, and chatbot output. Load testing confirmed the platform's stability up to 40 concurrent users* without performance degradation, ensuring its suitability for mid-scale deployment in educational and social platforms. These results are compiled in Table II.

Table II: OCR and System Performance Evaluation

Metric	Value	Description
Character Error Rate (CER)	4.8%	Avg. text extraction error from memes
Word Error Rate (WER)	6.3%	Avg. word-level OCR inaccuracy
End-to-End System Latency	1.5 seconds	Time from input to chatbot response
Concurrent User Support	40 users	Max load handled without system slowdown

D. Comparative Analysis

Compared to traditional cyberbullying detection tools that rely on keyword filtering or text-only inputs, the proposed system significantly outperforms in terms of detection accuracy, multimodal input handling, and user support. By integrating image-based text extraction, emoji interpretation, and dense AI-driven counselling, it offers a holistic approach to online abuse mitigation. The inclusion of specific bullying-type classification and multilingual adaptability further enhances the system's real-world applicability.

VI. CONCLUSION AND FUTURE SCOPE

This paper presented a Multimodal Cyberbullying Detection and Counselling System that addresses the rising concern of online abuse across social media platforms. By combining OCR-based image text extraction, emoji interpretation, multilingual text analysis, and a HateBERT-powered classification model, the system accurately detects cyberbullying in various formats including comments, memes, and emoji-rich messages. In addition to classifying bullying as age-based, gender-based, religion-based, and more, the system's multilingual capabilities enable it to effectively process code-mixed and non-English content and it offers immediate emotional support through a densely trained AI counselling chatbot.

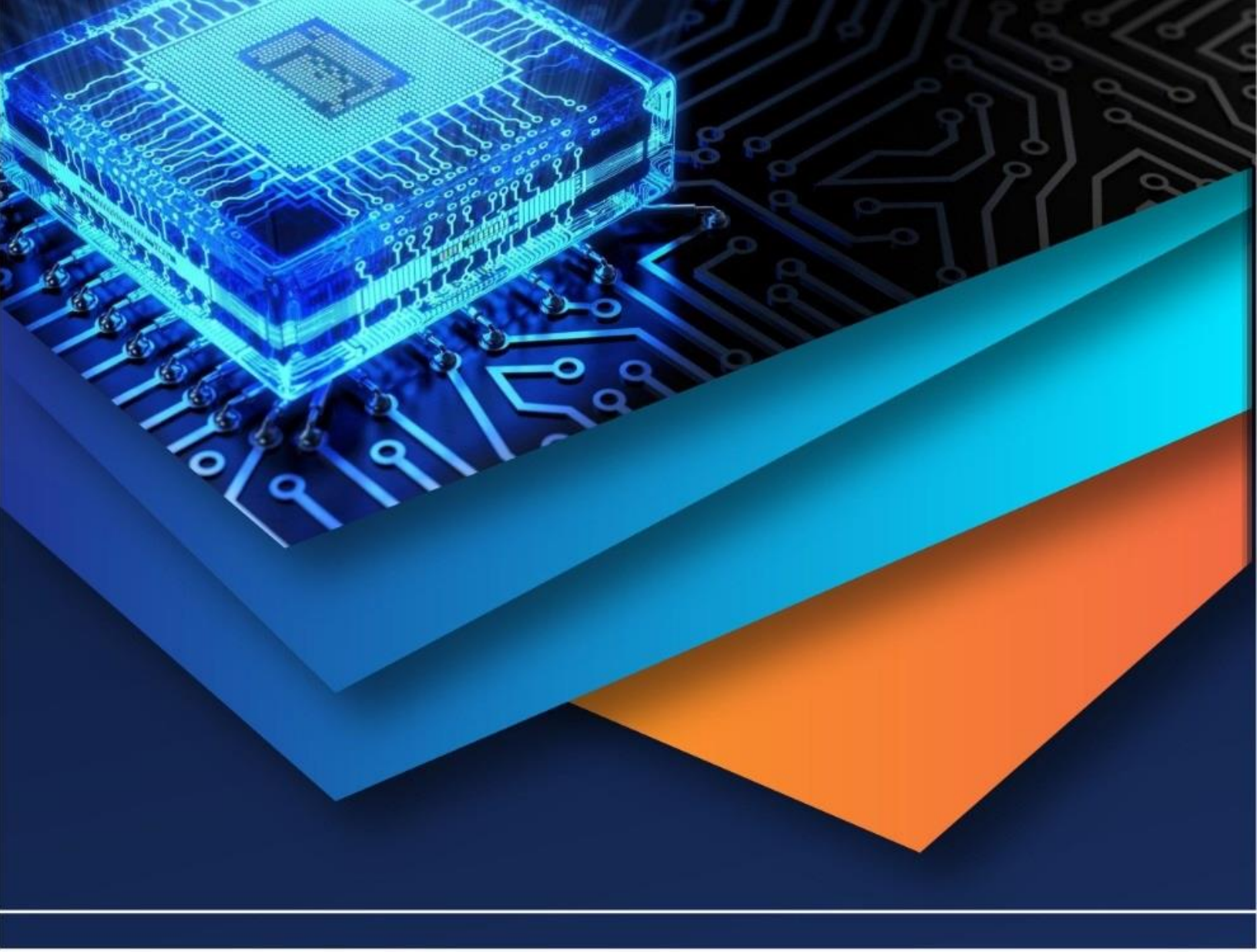
While the current implementation performs well across multiple dimensions, several future enhancements are identified to further strengthen the system:

1) Sarcasm and Coded Language Handling: Integration of models trained on sarcastic and contextually disguised bullying content will enhance sensitivity to implicit abuse.
2) Social Media API Integration: Real-time monitoring via direct platform APIs (e.g., Instagram, Twitter) would enable dynamic detection and user protection.
3) Adaptive Feedback Learning: Incorporating feedback loops from users to retrain the model can improve detection accuracy over time.
4) Cloud-Based Deployment: Hosting the system on scalable cloud infrastructure will allow deployment across institutions, public forums, and organizations.

In conclusion, the proposed system sets a comprehensive framework for intelligent, multilingual, and cross-modal cyberbullying detection with embedded emotional support. Its real-world adaptability positions it as a promising tool for promoting digital safety and psychological well-being in increasingly diverse online communities.

REFERENCES

[1] Alabdulwahab, A., Haq, M. A., &Alshehri, M. (2023). Cyberbullying Detection using Machine Learning and Deep Learning. International Journal of Advanced Computer Science and Applications, 14(10).
[2] Almomani, A., Nahar, K., Alauthman, M., Al-Betar, M. A., Yaseen, Q., & Gupta, B. B. (2024). Image Cyberbullying Detection and Recognition Using Transfer Deep Machine Learning. International Journal of Cognitive Computing in Engineering, 5, 14–26.
[3] Maity, K., Saha, S., & Bhattacharyya, P. (2023). Emoji, Sentiment, and Emotion Aided Cyberbullying Detection in Hinglish. IEEE Transactions on Computational Social Systems.
[4] Nikitha, G. S., Shenoyy, A., Chaturya, K., Latha, J. C., & Janani Shree, M. (2024). Detection of Cyberbullying Using NLP and Machine Learning in Social Networks for Bi-Language. International Journal of Scientific Research & Engineering Trends, 10(1), 128–134.
[5] Satya Narayana, G., Susmitha, V., Nagarani, J., Chinnarao, M., & Lavanya, P. (2024). Detection of Cyberbullying on Social Media Using Machine Learning Algorithms. International Journal of Novel Research and Development, 9(3), 45–52.
[6] Tahmid, F. I., Akbar, F., & Rahman, A. (2024). BulliShield: A Smart Cyberbullying Detection and Reporting System. IEEE Women in Data Science Conference, 198–203.
[7] Mahmud, T., Ptaszynski, M., & Masui, F. (2024). Exhaustive Study into Machine Learning and Deep Learning Methods for Multilingual Cyberbullying Detection in Bangla and Chittagonian Texts. Electronics, 13(9), 1677.
[8] Roy, P., & Mali, F. U. (2022). Cyberbullying Detection Using Deep Transfer Learning. Complex & Intelligent Systems, 8(6), 5449–5467.
[9] Rosa, H., Ribeiro, E., Ferreira, P. C., Carvalho, J. P., &Figueira, Á. (2023). Multimodal Cyberbullying Detection on Social Media Using Fusion of Text, Image, and Metadata. IEEE Access, 11, 5732–5744.
[10] Van Hee, C., Lefever, E.,&Hoste, V. (2023). Detection and Fine-Grained Classification of Cyberbullying Events. Natural Language Engineering, 29(2), 269–299.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY