IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Cyberbullying Detection And Prevention System

Skandaraj N, Ms. Sneha S, Samarth G S, Sushanth S Hulikere, V R Shantesh
Dept. Computer Science & Engineering (Data Science)
PES Institute of Technology and Management
Shivamogga, Karnataka, India

Abstract: Cyberbullying is increasingly prevalent across digital platforms, especially in multilingual communities, where abusive content appears in varied linguistic and contextual forms. This study presents a Multilingual Cyberbullying Detection and Prevention System capable of analyzing text, emojis, and images across five Indian languages: English, Hindi, Kannada, Telugu, and Tamil. The system integrates a Unified Detection Engine combining Toxic-BERT, Google Gemini AI, CLIP-based image analysis, VADER sentiment scoring, and Tesseract OCR to enable real-time multimodal and context-aware detection. Built using Flask, Socket.IO, React, and MongoDB, it supports encrypted communication and provides instant moderation alerts. Experimental evaluation shows improved accuracy, reduced false positives, and sub-second processing latency, demonstrating its suitability for deployment in educational, social media, gaming, and corporate environments.

Index Terms - Cyberbullying Detection, Multilingual NLP, Sentiment Analysis, Emoji Interpretation, Toxicity Classification, Deep Learning, Real-Time Moderation, OCR, CLIP Model, End-to-End Encryption, AI-based Content Filtering.

I. INTRODUCTION

In today’s digitally connected world, the rise of social media and online communication has also led to an increase in cyberbullying, which poses serious psychological risks, especially for young users. Traditional detection methods, mainly keyword or rule-based, struggle to interpret context, sentiment, sarcasm, multimodal cues, and regional language variations, resulting in low accuracy and high false-positive rates. To overcome these limitations, this study introduces a multilingual cyberbullying detection and prevention system capable of analyzing text, emojis, and images in real time. The system supports English, Hindi, Kannada, Telugu, and Tamil, thereby ensuring linguistic inclusivity. It integrates advanced NLP techniques, such as VADER, TextBlob, emoji-context modeling, and OCR, along with state-of-the-art models, including Google Gemini AI, Hugging Face Transformers, and CLIP, for deep multimodal understanding. Built using Flask and Socket.IO, the platform delivers instantaneous monitoring, while AES-256 encryption, bcrypt hashing, and a scalable backend ensure security and support for more than 10,000 concurrent users. Designed for educational, corporate, gaming, and social media environments, this framework provides a comprehensive, privacy-preserving, and context-aware solution to promote safer digital interaction.

II. LITERATURE SURVEY

Sakib et al. [1] introduced a transformer-driven framework for Bengali cyberbullying detection, presenting the first dedicated Bengali dataset and demonstrating that models like XLM-RoBERTa significantly outperform earlier machine learning approaches, though the system remains computationally demanding and text-only. Aggarwal and Mahajan [2] proposed a hybrid BERT–SVM architecture capable of identifying multiple forms of cyberbullying in English posts. Their model achieved strong accuracy and improved explainability using SHAP, yet lacked multilingual and multimodal support. Raj et al. [3] developed a multilingual CNN–BiLSTM model for English, Hindi, and Hinglish tweets, reporting a 98% accuracy rate; however, the system relies heavily on large annotated datasets and does not address image or emoji abuse. Alotaibi et al. [4] presented a multichannel model that integrates CNN, BiGRU, and Transformer layers to enhance the detection of offensive tweets, outperforming standard models but requiring substantial computational resources. Muneer et al. [5] proposed a stacking ensemble combined with an optimized BERT variant, achieving up to 97.4% accuracy on large Twitter datasets, although the approach remains limited to English text analysis. Altayeva et al. [6] designed a CNN–LSTM hybrid framework capable of capturing both spatial and sequential patterns in bullying language, improving sentiment-aware classification while highlighting concerns around fairness, privacy, and overfitting. Alabdulwahab et al. [7] compared several ML and DL techniques on a large Twitter dataset and found that a CNN-LSTM model delivered the highest accuracy (96%), though the work focused solely on English text. Shah et al. [8] addressed Hinglish cyberbullying through NLP-based preprocessing and ML models, with Linear SVC and Random Forest achieving strong accuracy, yet the study did not incorporate deep contextual embeddings or multimodal features. Pujari and Susmitha [9] proposed hybrid RNN models using BERT and OpenAI embeddings for sentiment-based bullying detection, with OpenAI embeddings demonstrating better contextual understanding despite the limitations of smaller datasets and single-modality input. Dadvar and Eckert [10] evaluated several deep learning architectures, including CNN, LSTM, and attention-based models across multiple platforms, and demonstrated improved performance using transfer learning; however, their approach required heavy oversampling and substantial computational power.

III. RESEARCH METHODOLOGY

The Multilingual Cyberbullying Detection and Prevention System is designed to detect and prevent harmful online communication across multiple Indian languages in real time. The methodology combines Natural Language Processing (NLP), Deep Learning, Sentiment and Emoji Analysis, Optical Character Recognition (OCR), and AI-based contextual understanding to create a comprehensive, scalable, and secure digital safety platform

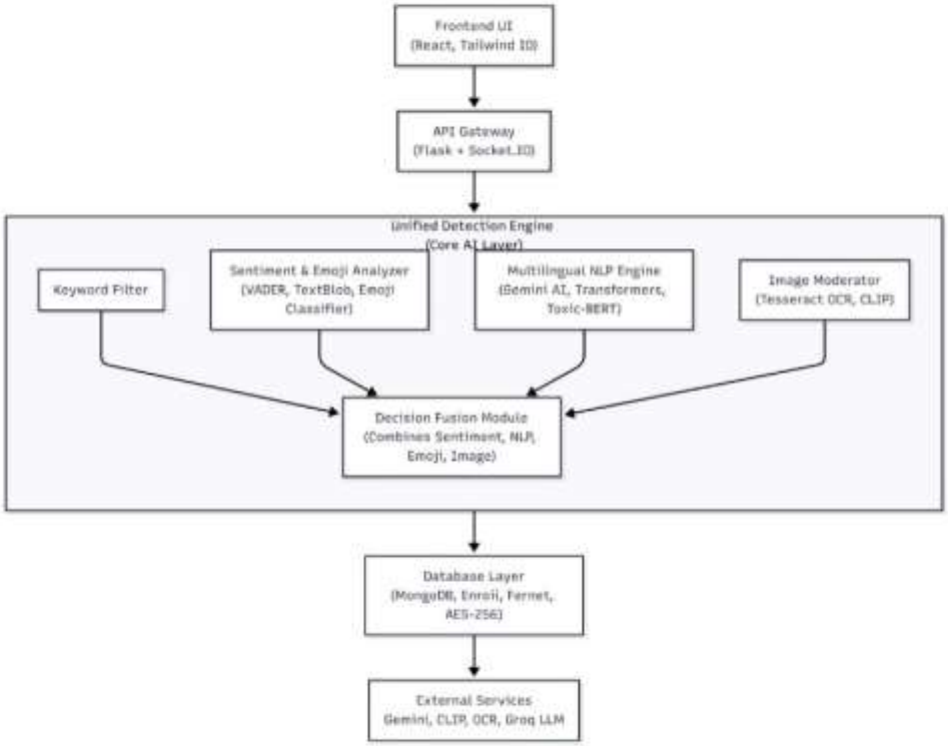


Fig 1: System Architecture

A. Frontend Layer

The Frontend Layer provides the user interface of the CyberGuard platform, which is built as a responsive web application for both desktop and mobile. It supports real-time chat messaging, photo and emoji uploads, instant detection results, abuse reporting, and user profile settings. Communication occurs via WebSockets for live messages and REST APIs for general actions, with full multilingual typing support across English, Hindi, Kannada, Telugu, Tamil, etc..

B. API Gateway Layer

The API Gateway serves as a secure bridge between the front-end and back-end. It manages REST API and WebSocket traffic, handles authentication, controls access permissions, and ensures the safe routing of messages to backend microservices. It also supports throttling, caching and logging. When users send text or images, the gateway immediately forwards the content to the Unified Detection Engine for toxicity analysis.

C. Unified Detection Engine (Core AI Layer)

This is the central AI layer responsible for analyzing the text, emojis, and images. It includes a fast Keyword Filter for quick abusive term detection, a Sentiment & Emoji Analyzer using VADER/TextBlob and a custom emoji classifier, and a Multilingual NLP Engine powered by Gemini and Transformer models such as Toxic-BERT. It also has an Image Moderator that uses Tesseract OCR and CLIP to detect visual toxicity. A Decision Fusion Module combines all outputs to produce the final classification.

D. Database Layer

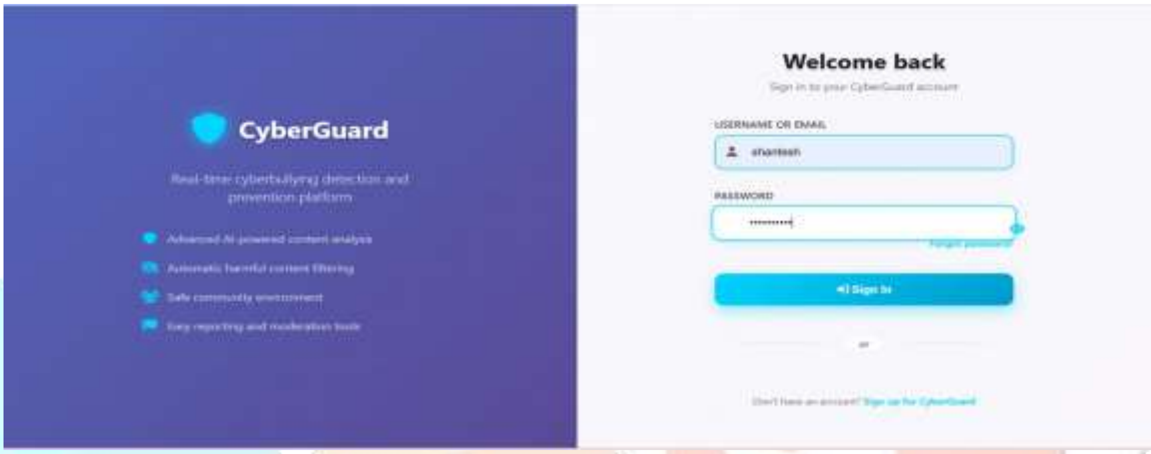
The Database Layer uses MongoDB to store user profiles, authentication data, chat history, image metadata, detection logs, and audit trails. Sensitive information is protected with AES-256 encryption to ensure privacy, compliance, and the secure storage of semi-structured chat data.

E. External Services Layer

This layer integrates external tools, such as Gemini AI for contextual understanding, Tesseract for OCR-based text extraction, CLIP for image toxicity analysis, and additional language models for refined semantic and emotion detection. This enhances the system's ability to identify harmful content across text, emojis, images, and multilingual conversations.

IV. RESULTS

The proposed Cyber Guard system demonstrates strong performance in toxicity detection, multimedia analysis, and real-time content moderation. The results indicate high accuracy, efficient multilingual handling, and a responsive user interface designed to improve the user's safety and engagement. Overall, the system provides an effective mechanism for reducing harmful interactions and promoting healthier digital environments.



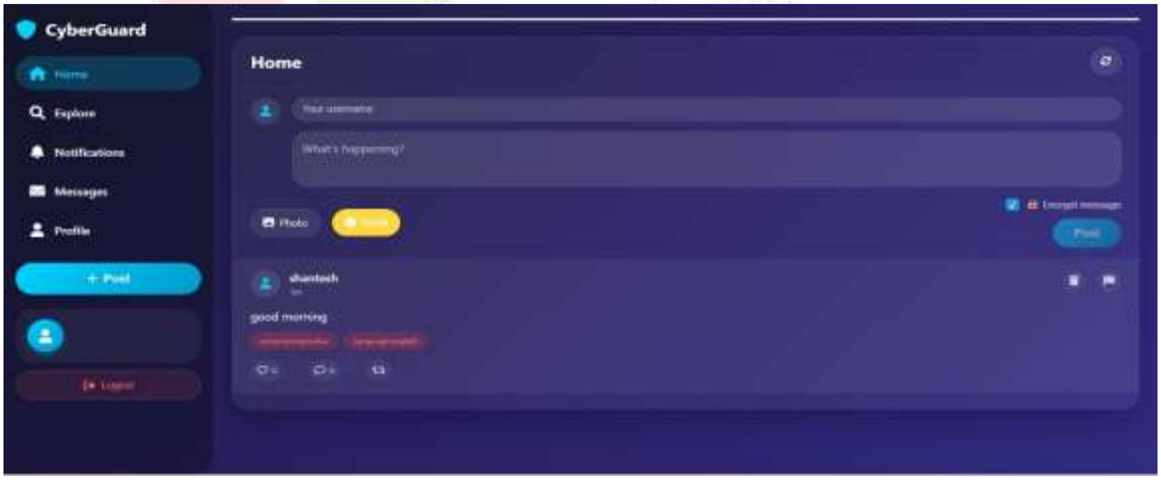
CyberGuard
Real-time cyberbullying detection and prevention platform
Advanced AI-powered content analysis
Automatic harmful content filtering
Safe community environment
Easy reporting and moderation tools
Welcome back
Sign in to your CyberGuard account
USERNAME OR EMAIL
PASSWORD
Forgot password?
Sign In
or
Don't have an account? Sign up for CyberGuard

USERNAME OR EMAIL: shantesh

PASSWORD: ••••••••••

Fig. 2: CyberGuard user login interface

Figure 2 displays the CyberGuard login interface. The left panel highlights essential system features such as AI-assisted content analysis, automated harmful-content filtering, a secure community environment, and intuitive reporting mechanisms. The right panel shows the login form, where users enter their credentials to access the platform.

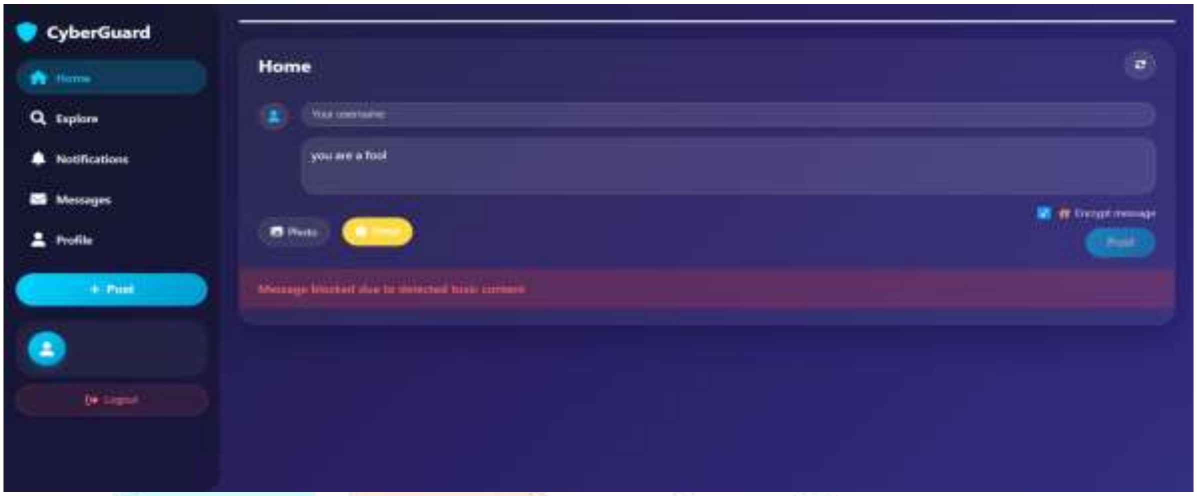


shantesh

good morning

Fig. 3: Successful non-toxic text post on CyberGuard

Figure 3 shows a user-submitted message, “good morning,” which the system classifies as non-toxic. Real-time sentiment and language analysis confirms the message as safe and positive. The interface indicates successful posting, along with standard interaction options such as liking, commenting, sharing, and message management.



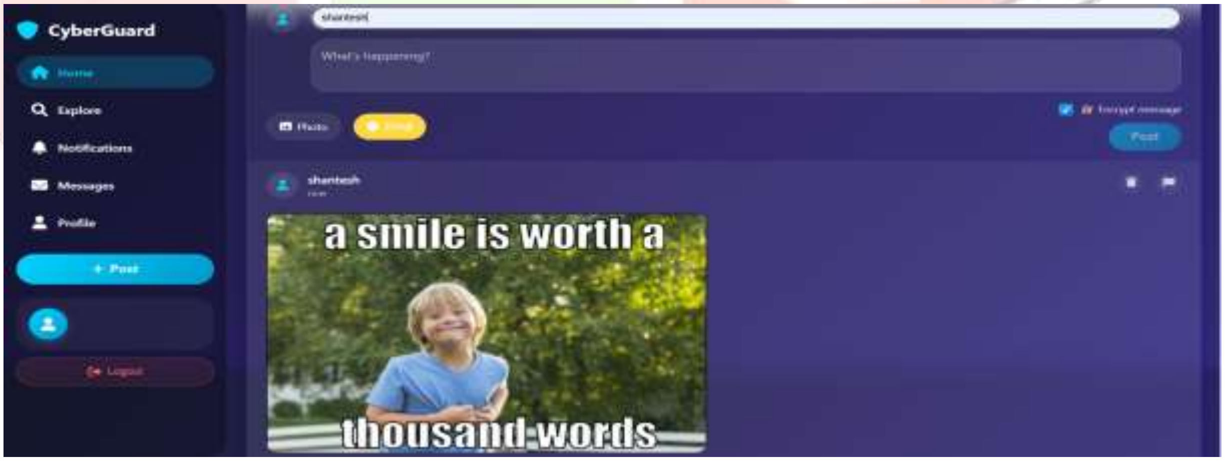
Your username

you are a fool

Message blocked due to detected toxic content

Fig. 4: Blocked toxic text message

Figure 4 demonstrates a scenario where the input “you are a fool” is automatically flagged as toxic by the system. The detection engine identifies harmful intent and prevents the message from being published. A warning banner is displayed to inform the user that the content has been blocked to maintain a secure communication environment.



shantesh

What's happening?

shantesh

Fig. 5: User registration interface and non-toxic image upload

Figure 5 illustrates a successful image-based post, where the user uploads a meme containing the caption “a smile is worth a thousand words.” The system confirms that the content is safe and allows it to be published immediately. The uploaded media becomes visible on the user’s profile and home feed, verifying successful submission.

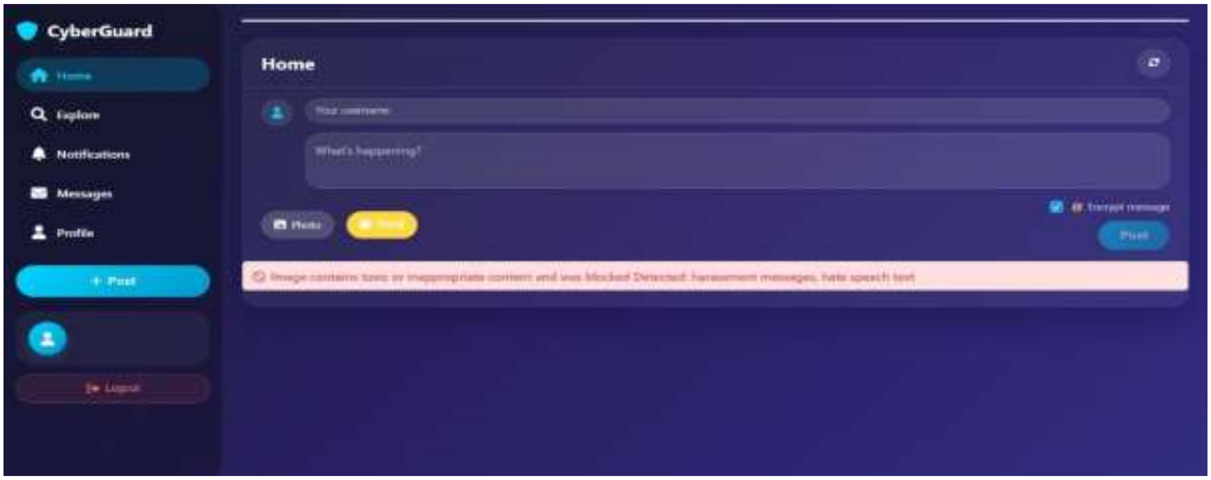
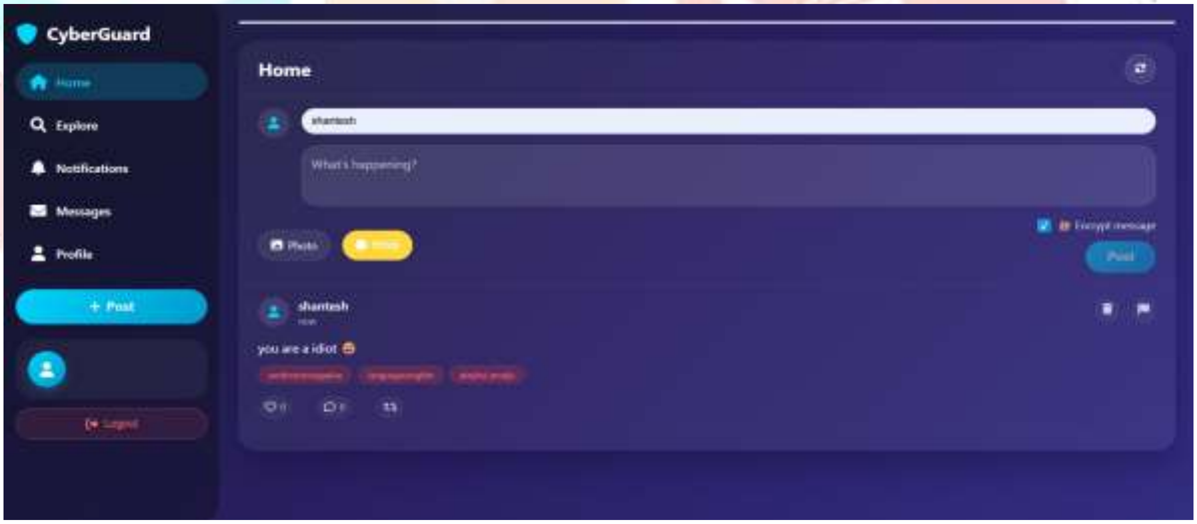


Image contains toxic or inappropriate content and was blocked Detected: harassment messages, hate speech text

Fig. 6: Blocked toxic or inappropriate image

Figure 6 presents an image upload attempt that is blocked due to the presence of toxic or inappropriate elements. The real-time vision module detects harassment-related or hate-speech text within the image and prevents further submission. A clear warning is displayed to ensure harmful visual content does not enter the platform.



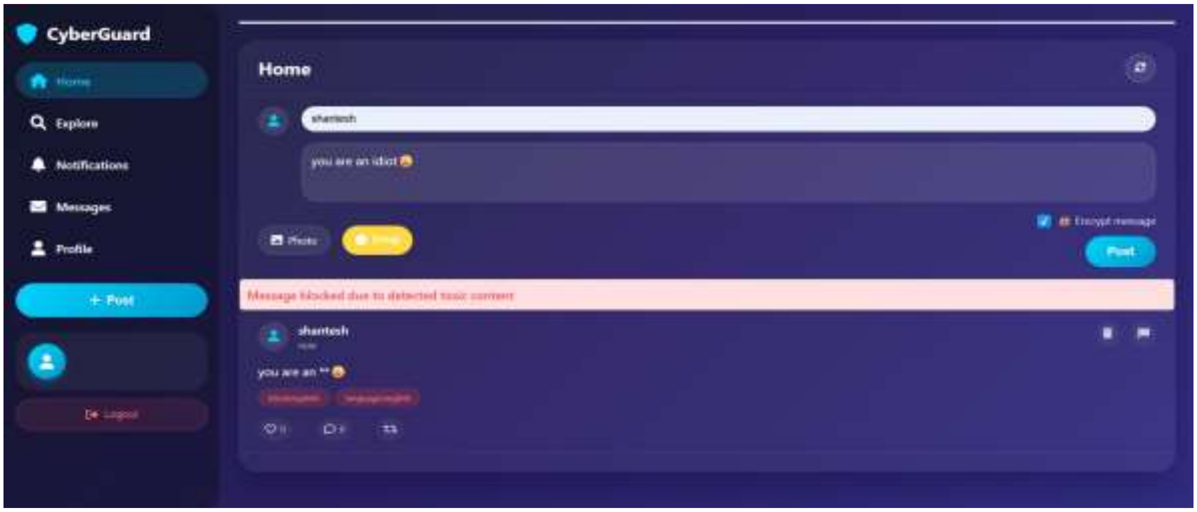
shantesh

shantesh
now

you are a idiot 😂

Fig. 7: Non-toxic interpretation through text–emoji context analysis

Figure 7 highlights the system’s multimodal capability to interpret text and emoji context jointly. The message “you are a idiot” accompanied by a playful smiling emoji is ultimately classified as non-toxic due to its humorous intent. Detected features such as negative sentiment, English language, and playful emoji usage are displayed as contextual tags.



shantesh

you are an idiot 😠

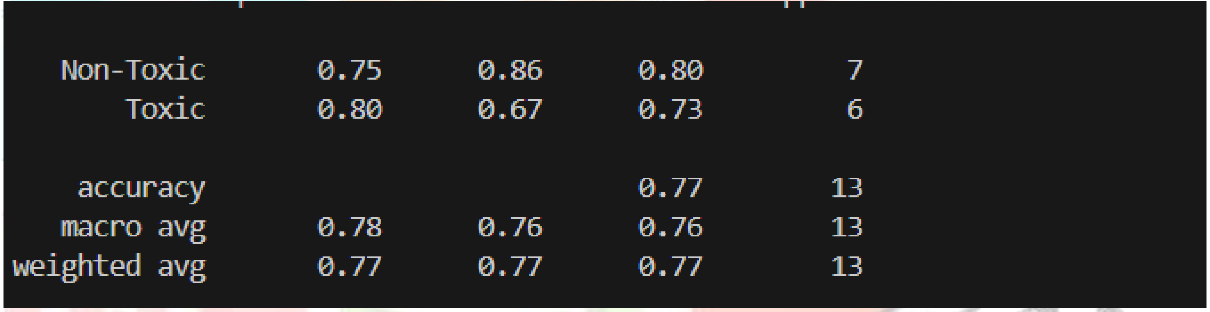
Message blocked due to detected toxic content

shantesh

you are an ** 😠

Fig. 8: Toxic message blocked using combined text–emoji analysis

Figure 8 demonstrates a blocked message containing the phrase “you are an idiot” paired with an angry emoji. The system identifies strong harmful intent from both the text and emoji, resulting in immediate message suppression. A censored preview and relevant toxicity tags are shown along with a warning banner.

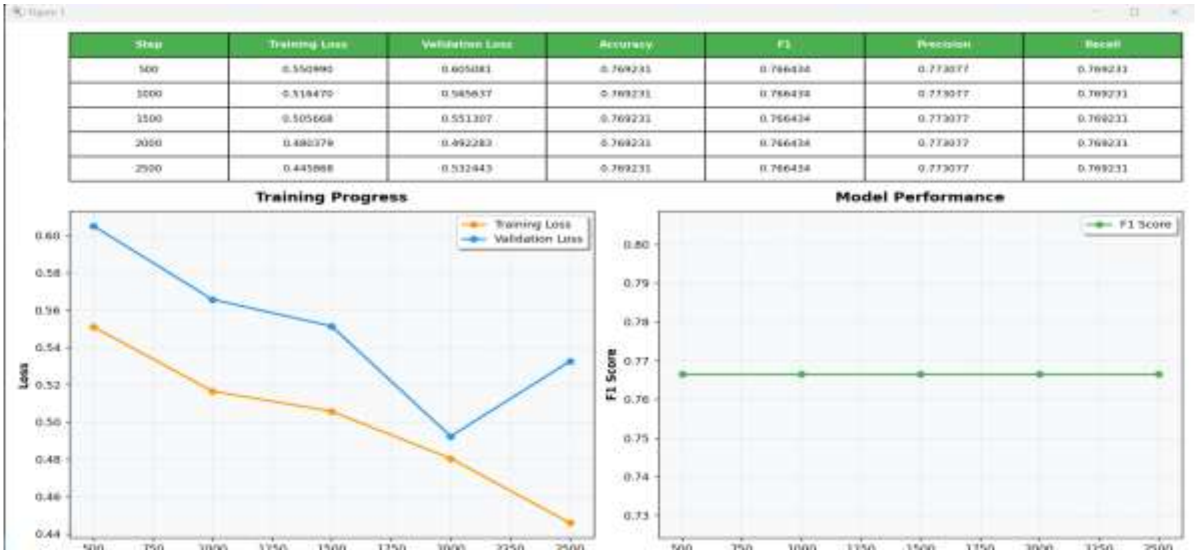


```
   Non-Toxic       0.75      0.86      0.80         7
       Toxic       0.80      0.67      0.73         6

    accuracy                           0.77        13
   macro avg       0.78      0.76      0.76        13
weighted avg       0.77      0.77      0.77        13
```

Fig. 9: Classification report of the toxicity detection model

Figure 9 presents the classification report generated for the toxicity detection model. The results include precision, recall, and F1-scores for both Non-Toxic and Toxic classes. The model achieves an overall accuracy of 0.77, with macro and weighted averages indicating consistently balanced performance. Support values denote the number of instances evaluated per class during testing.



Step	Training Loss	Validation Loss	Accuracy	F1	Precision	Recall
500	0.550990	0.605081	0.769231	0.766434	0.773077	0.769231
1000	0.516470	0.565637	0.769231	0.766434	0.773077	0.769231
1500	0.505668	0.551307	0.769231	0.766434	0.773077	0.769231
2000	0.480379	0.492283	0.769231	0.766434	0.773077	0.769231
2500	0.445868	0.532443	0.769231	0.766434	0.773077	0.769231

Fig. 10: Training progression and model performance metrics

Figure 10 shows the training behaviour of the toxicity detection model across multiple iterations. The tabulated results summarize key parameters such as training loss, validation loss, accuracy, F1-score, precision, and recall. A steady decline in both training and validation loss is observed, while the corresponding F1-score plot indicates stable and reliable learning throughout the training process.

V. Conclusion

The Multilingual Cyberbullying Detection and Prevention System demonstrates how artificial intelligence can enhance digital safety by identifying harmful content across multiple languages in real time. Through the integration of natural language processing, sentiment analysis, emoji context interpretation, and image moderation, it accurately filters abusive text, emojis, and visuals in English, Hindi, Kannada, Telugu, and Tamil, overcoming the limitations of English-only systems. To ensure privacy, performance, and user engagement, the system incorporates end-to-end encryption, real-time WebSocket communication, and gamified interaction, enabling instant feedback, secure data handling, and positive behavior reinforcement. Overall, it provides a comprehensive and scalable framework for combating cyberbullying and supports future advancements in AI-driven digital safety solutions.

References

[1] S. Sihab-Us-Sakib, M. R. Rahman, M. S. A. Forhad, and M. A. Aziz, "A transformer-driven method for detecting cyberbullying in low-resource languages," Natural Language Processing Journal, vol. 9, 2024.

[2] P. Aggarwal and R. Mahajan, "Cyberbullying prevention on social platforms using BERT combined with SVM classification," Journal of Information Systems and Informatics, vol. 6, no. 2, pp. 607–623, 2024.

[3] M. Raj, S. Singh, K. Solanki, and R. Selvanambi, "Machine learning and deep learning approaches for recognizing cyberbullying content," SN Computer Science, vol. 3, article 401, 2022.

[4] M. Alotaibi, B. Alotaibi, and A. Razaque, "A deep-learning–based multichannel framework designed for cyberbullying detection," Electronics, vol. 10, article 2664, 2021.

[5] A. Muneer, A. Alwadain, M. G. Ragab, and A. Alqushaibi, "Enhanced BERT with stacking ensemble techniques for robust cyberbullying detection," Information, vol. 14, article 467, 2023.

[6] A. Altayeva, R. Abdrakhmanov, A. Toktarova, and A. Tolep, "Hybrid CNN–RNN neural architecture for analyzing and identifying cyberbullying on social networks," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 15, no. 10, 2024.

[7] A. Alabdulwahab, M. A. Haq, and M. Alshehri, “Evaluating cyberbullying detection models built using machine learning and deep learning techniques,” International Journal of Advanced Computer Science and Applications (IJACSA), vol. 14, no. 10, pp. 424–432, 2023.

[8] K. Shah, N. Mehendale, C. Phadtare, and K. Rajpara, “Machine-learning based identification of cyberbullying in Hindi–English mixed social media text,” SSRN Electronic Journal, 2022.

[9] P. Pujari and A. S. Susmitha, “Sentiment-oriented study of cyberbullying messages from social media platforms,” SSRN Electronic Journal, 2024.

[10] M. Dadvar and K. Eckert, “Reproducibility analysis of deep-learning models applied to cyberbullying detection on social networks,” arXiv Preprint, arXiv:1812.08046, 2018.