Grenze International Journal of Engineering and Technology, June Issue

Multimodal Cyberbullying Detection using Text, Emojis, and Images

Pulivarthi Sandhya[1], Atluri Aasritha[2], Golve Anusha[3], Pannala Renuka[4] and D. Rajeswara Rao[5]
[1-4]Department of Computer Science and Engineering, Velagapudi Ramakrishna Siddhartha Engineering College, Vijayawada, India
Email: sandhyapulivarthi6@gmail.com, aasrithaatluri@gmail.com, golveanusha@gmail.com, pannalarenuka13@gmail.com
[5]Professor & Head, Department of Computer Science and Engineering, Velagapudi Ramakrishna Siddhartha Engineering College, Vijayawada, India
Email: hodcse@vrsec.ac.in

***Abstract*—Cyberbullying has progressed from simple text and is now often shown through images, screenshots, and emojis, complicating detection. Conventional text-based models do not effectively harness these multimodal signals, leading to diminished precision in practical situations. To overcome this limitation, this paper proposes a multimodal cyberbullying detection system that combines Optical Character Recognition (OCR), deep learning text analysis, and visual feature extraction. The system utilizes OCR to extract text from images and processes it with DeBERTa embeddings paired with LSTM to capture contextual and sequential patterns. Moreover, emoji and visual indicators are examined through a CLIP-based feature extractor. A fusion approach integrates text and visual representations to categorize the input as toxic or non-toxic. The model attains impressive results with a validation accuracy of 94%, showcasing its capability in identifying cyberbullying through various modalities. The system is implemented through a Gradio-powered web interface, allowing for immediate evaluation of text and visuals.**

***Index Terms*— Cyberbullying Detection, Multimodal Learning, DeBERTa, LSTM, OCR, CLIP.**

I. INTRODUCTION

The swift expansion of social media sites and digital communication tools has notably heightened the occurrence of cyberbullying. In contrast to conventional text-based bullying, contemporary cyberbullying frequently involves a mix of text, images, screenshots, and emojis, resulting in increased complexity and making it harder to identify [5]. Harmful material often surfaces in images or is expressed indirectly through emojis and visual components, posing difficulties for conventional detection systems to identify correctly [1]. Consequently, numerous cases of cyberbullying remain undetected, presenting significant emotional and psychological dangers to people, particularly teenagers. The proposed system enhances the accuracy and robustness of detecting cyberbullying by integrating textual and visual representations using a fusion method [3]. Furthermore, a user-friendly web interface is created with Gradio, enabling users to enter text or upload images for analysis in real-time. This method improves the system's capability to identify both overt and subtle types of cyberbullying, making it more appropriate for practical use.

Grenze ID: 01.GIJET.12.2.185_26

A. Scope and Novelty

The scope of this work is to develop a multimodal framework for detecting cyberbullying that can analyze text, images, and emoji content typically found on social media and online communication channels. Conventional methods for detecting cyberbullying mainly concentrate on text and frequently overlook harmful content that is expressed through images, screenshots, and emojis. The suggested framework tackles this issue by combining various modalities to enhance the identification of both explicit and implicit instances of cyberbullying.

The novelty of the proposed framework lies in the integration of DeBERTa for contextual language comprehension, LSTM for learning sequential features, OCR for obtaining textual data from images and screenshots, and CLIP for identifying visual and emoji-related features. Current methods typically emphasize either text-only analysis or a restricted blend of text and images, while the suggested framework concurrently handles text, emojis, and image-based material in an integrated structure. Additionally, the multimodal fusion approach allows the system to recognize contextual, emotional, and visual signals that traditional cyberbullying detection techniques might miss, thus enhancing the accuracy of harmful content detection across various communication types.

B. Key Contributions

The key contributions of our work are as follows:

- Recognition of shortcomings in current cyberbullying detection methods, especially their failure to address multimodal content including images, emojis, and screenshots.
- Creation of a multimodal hybrid model combining DeBERTa and LSTM for text analysis, along with OCR and CLIP-based methods for interpreting visuals and emojis.
- Creation of a system able to grasp contextual significance, emotional nuance, and underlying intent in both text and image inputs.
- Creation of an intuitive Gradio interface enabling real-time detection via text entries and image uploads.
- Assessment and enhancement of the model to guarantee consistent performance across various online communication formats, such as social media updates, conversations, and multimedia materials.

II. Related Work

Ahmed et al. [4] introduced a multimodal method for detecting cyberbullying that emphasizes meme-oriented content from social media platforms. Their model integrates textual and visual characteristics through deep learning methods to detect harmful material in memes. While the method enhances detection precision in image-related cyberbullying, it struggles with understanding more profound contextual connections in text.

Kiela et al. [5] launched the Hateful Memes Challenge, aimed at identifying hate speech in multimodal meme datasets through the integration of image and text elements. The research emphasizes the necessity of combining visual and textual signals for precise categorization. Nonetheless, the model necessitates extensive datasets and computational power, rendering it less efficient for immediate applications.

Maity et al. [6] introduced a multimodal multitask frame-work that simultaneously learns from text and image features to identify cyberbullying in memes. Their method enhances performance by utilizing common representations across tasks. Nevertheless, the system might have difficulty grasping implicit context and emotional subtleties.

Jha et al. [7] created a multimodal framework based on explanations for detecting cyberbullying, which forecasts harmful content while also offering clarity regarding the decisions made. Though this enhances transparency, it also raises model complexity, complicating deployment.

Singh et al. [8] developed a multimodal system for detecting cyberbullying with severity assessment, integrating both textual and visual attributes to evaluate the degree of harmful material. While the system improves detection ability, it still struggles with efficiently managing sequential dependencies in text data.

Mohiuddin et al. [9] introduced a multimodal system based on deep learning that utilizes OCR to extract text from images and analyze it alongside visual characteristics. This method enhances the identification of concealed text in screenshots and memes; however, it might be affected by image quality and OCR precision.

Roy et al. [10] introduced a classification model that combines image and text data for detecting cyberbullying. Their method enhances precision in multimodal settings yet lacks sophisticated contextual comprehension capabilities.

Barbieri et al. [11] introduced an emoji embedding model that captures the emotional significance expressed through emojis in digital interactions. Their method aids in comprehending sentiment and contextual signals conveyed through emojis, commonly utilized in cyberbullying situations.

Recent studies have demonstrated the effectiveness of transformer-based multimodal models in identifying cyberbullying and harmful content. Comprehensive vision-language models and adaptable fusion methods have shown improved capability in identifying nuanced abusive language present in text, images, and emojis. However, many existing techniques either require substantial computational resources or fail to effectively represent sequential contextual connections. These limitations inspire the development of the proposed DeBERTa and LSTM multimodal system, which unifies contextual understanding, sequential processing, and efficient execution.

TABLE I. COMPARISON BETWEEN RECENT CYBERBULLYING APPROACHES AND PROPOSED MODEL

Title	Model	Datasets	Advantage	Disadvantage
Ahmed et al., 2023	Multimodal DL	Meme Dataset	Combines text and image features for detecting meme-based cyberbullying	Weak contextual understanding in textual content
Kiela et al., 2020	Multimodal Trans-former	Hateful Memes Dataset	Integrates image and text for accurate hate detection	Requires large dataset and high computational cost
Maity et al., 2022	Multitask Multimodal Model	Meme Dataset	Learns shared representations across text and image tasks	Struggles with implicit context and emotional cues
Jha et al., 2024	Explainable Multimodal Model	Social Media Memes	Provides interpretability along with detection	High model complexity and difficult deployment
Singh et al., 2025	Multimodal DL + Severity Analysis	Social Media Dataset	Classifies level of cyberbullying severity	Limited sequential understanding of text
Mohiuddin et al., 2025	OCR + Deep Learning	Image + Text Dataset	Detects hidden text in images using OCR	Sensitive to im-age quality and OCR errors
Roy et al., 2025	Multimodal Classification Model	Image + Text Dataset	Improves accuracy using combined modalities	Lacks strong contextual understanding
Barbieri et al., 2018	Emoji Embedding Model	Emoji Dataset	Captures emotional meaning of emojis	Cannot work independently without text/image context

III. PROPOSED METHODOLOGY

The primary objective of the proposed framework is to improve the accuracy and reliability of cyberbullying detection across multimodal inputs. The process flow diagram presented in Figure 1 graphically demonstrates the sequential steps within the operation of the system, ensuring clear comprehension of the overall process. The suggested framework for detecting cyberbullying includes various interconnected elements that collaborate to guarantee precise classification of multimodal inputs. The process starts with data preprocessing, which entails cleaning textual information by eliminating noise, converting to lowercase, tokenizing, and standardizing sentence lengths via padding or truncation. For visual inputs like chat screenshots and memes, Optical Character Recognition (OCR) is utilized to retrieve embedded text, which is subsequently handled in the same way as textual inputs. The system utilizes DeBERTa, a transformer architecture, to produce contextual embeddings that grasp profound semantic connections among words. These embeddings are subsequently fed into an LSTM network, which captures sequential and latent features, allowing the model to comprehend both the sequence and deeper significance of the content. Moreover, a CLIP-based model is utilized to extract emoji and visual elements to grasp the emotional and contextual signals found in images.

The combined multimodal features are forwarded through a dense classification layer utilizing sigmoid or softmax activation to determine if the content is toxic or non-toxic, along with a confidence score. Additionally, a user interface is created with Gradio, enabling users to enter text or upload images to assess the model instantly. This comprehensive system successfully combines cutting-edge natural language processing, deep learning, and multimodal analysis methods to ensure dependable detection of cyberbullying. The modular structure also guarantees adaptability for future improvements, like expanding to multi-class classification or adding more modalities.

A. Gathering Dataset

To guarantee efficient processing and quicker experimentation, a subset of 19,531 samples is utilized for model training and evaluation. The dataset mainly consists of text data, with image inputs processed independently at runtime through OCR and emoji extraction methods. This integration fosters the creation of a strong multimodal model adept at addressing both text-based and visual cyberbullying material.

B. Data Preparation

The process starts with the preprocessing of input data sourced from different cyberbullying datasets, encompassing both direct text entries and text retrieved from images through OCR. The steps for preprocessing are outlined as follows:

- Text Sanitization: Elimination of undesirable elements like special symbols, hyperlinks, and excess whitespace from user-provided text and OCR-derived text.
- Lowercasing: Transforming all text into lowercase to guarantee consistency and decrease the size of the vocabulary.
- Tokenization: A transformer-based tokenizer is employed to segment text into significant tokens or sub-words, enhancing semantic comprehension.
- Padding and Truncation: Every sequence is modified to a predetermined length to guarantee uniform input for the DeBERTa-LSTM model

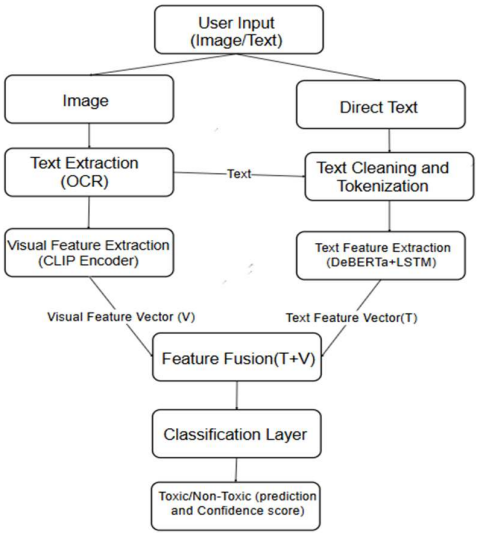

Figure 1. Process flow of proposed cyberbullying detection model

C. Text Encoding Using Transformer Tokenizer

Transformer-based models require numerical input. After preprocessing, a pretrained DeBERTa tokenizer encodes the text, generating attention masks and converting the text into token IDs. This ensures adherence to the transformer structure while maintaining semantic content.

D. Embedding Creation (DeBERTa)

The DeBERTa model takes the encoded text and utilizes it to generate contextual embeddings. These embeddings capture semantic meaning and contextual interactions between words, allowing the technology to detect subtle and implicit indicators of cyberbullying as shown in Algorithm 1.

Algorithm 1: Extract CLS Token Embedding from DeBERTa

1: Input: Encoded input E, pretrained DeBERTa model
2: Output: CLS embedding vector C
3: Disable gradient computation
4: Pass input IDs and attention masks to DeBERTa model
5: Obtain last hidden state from model output
6: Extract CLS token embedding C = hidden state[:, 0, :]
7: Return CLS embedding C

E. Feature Extraction (LSTM) and Model Training

LSTM is employed for the sequential examination of the tokens produced following the embedding creation process with DeBERTa. The DeBERTa model produces embeddings that are then fed into a Long Short-Term Memory (LSTM) network for sequential feature extraction and classification.

F. Module for Optical Character Recognition (OCR)

The system includes an OCR component to pull text from images like chat screenshots and memes. This enables the model to analyze text contained in images, facilitating the identification of cyberbullying material in addition to simple text inputs.

G. Extraction of Visual Features and Emojis (CLIP)

The system employs a CLIP-oriented model to derive visual and emoji-related characteristics from input images. Emojis found in images frequently communicate emotional context that may not be explicitly articulated in text. The CLIP model translates visual data into feature vectors, allowing the system to interpret emotional signals and contextual significance linked to the content. Algorithm 2 shows the process of extracting visual features using CLIP.

Algorithm 2: Extract Visual Features using CLIP

1: Input: Image I, pretrained CLIP model
2: Output: Feature vector F
3: Preprocess image I
4: Pass image through CLIP image encoder
5: Extract feature representation F
6: Return

H. Fusion of Multimodal Features

The system integrates textual attributes derived from DeBERTa-LSTM with visual and emoji attributes obtained through the CLIP model.

This combination allows the model to evaluate textual and visual data at the same time. The integrated feature vector enhances the system's capability to identify cyberbullying by including contextual, sequential, and emotional signals.

The combined representation is subsequently sent to a classification layer for the final prediction. Algorithm 3 shows the fusion of multimodal features.

Algorithm 3: Multimodal Feature Fusion

1: Input: Text features T , Visual features V
2: Output: Fused feature vector F
3: Concatenate T and V
4: Apply dense layer for feature combination
5: Return fused representation F

I. Layer for Classification

The combined feature vector is sent through a fully connected layer for classification purposes. A sigmoid or softmax activation function is used to produce probability scores that denote whether the input is toxic or not toxic

J. Experimental Setup

The proposed model was trained and evaluated using a dataset comprising 19,531 samples. The framework utilizes DeBERTa for contextual text representation, LSTM for capturing sequential features, OCR for extracting textual information from images and screenshots, and CLIP for identifying visual and emoji-related features. The performance of the proposed model was evaluated using standard classification metrics, including Accuracy, Precision, Recall, and F1-Score. These metrics were used to assess the effectiveness of the framework in identifying both toxic and non-toxic content across different input modalities.

IV. RESULT AND ANALYSIS

A. Performance Analysis

- Confusion Matrix: To evaluate the effectiveness of the proposed model, a confusion matrix is used. It provides a detailed comparison between actual and predicted classifications. Figure 2 shows that most non-toxic and toxic samples are correctly classified, while only a small number of misclassifications are observed. Out of 13,021 actual "Non-Toxic" samples, 12,300 were correctly identified as "Non-Toxic", while from 6,510 actual "Toxic" cases, 6,059 samples are correctly classified.

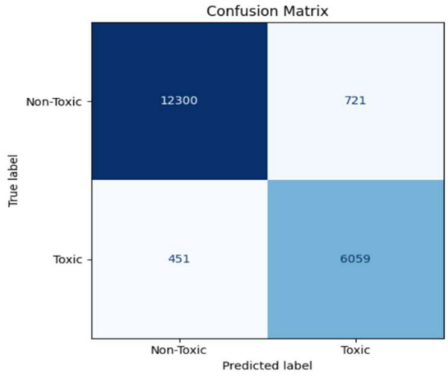

Figure 2: Confusion Matrix

B. Tabular Representation of Metrics

The system processes different types of inputs through multiple stages, as shown in TABLE II. The classification metrics indicate balanced performance across both classes. As shown in TABLE III, the model achieved a Precision of 0.96, Recall of 0.94, and F1-Score of 0.95 for the non-toxic class, while obtaining a Precision of 0.89, Recall of 0.93, and F1-Score of 0.91 for the Toxic class.

TABLE II. INPUT PROCESSING AND OUTPUT REPRESENTATION

Input type	Processing Technique	Output
Text	DeBERTa + LSTM	Toxic/non-toxic
Image	OCR + CLIP	Extracted features
Emoji	CLIP Model	Emotional Content
Combined Input	Feature Fusion	Final Prediction

TABLE III. CLASSIFICATION METRICS FOR CYBERBULLYING DETECTION

Class	Precision	Recall	F1-Score	Support
Non-Toxic	0.96	0.94	0.95	13021
Toxic	0.89	0.93	0.91	6510
Accuracy	0.94 (on 19531 samples)			
Macro Avg	0.92	0.93	0.93	19531
Weighted Avg	0.94	0.94	0.94	19531

The evaluation results indicate that the proposed approach performs effectively in distinguishing toxic and non-toxic content across multiple data modalities. An overall accuracy of 94% was achieved, accompanied by strong Precision, Recall, and F1-Score values for both classes. The weighted F1-Score of 0.94 reflects consistent classification performance and the model's ability to generalize across diverse samples. In addition, the incorporation of OCR and CLIP allows the framework to analyze information extracted from screenshots, images, and emoji-based interactions, enabling the identification of cyberbullying patterns that may not be captured through text analysis alone.

B. User Interface

The proposed multimodal framework detects cyberbullying effectively by combining contextual understanding using DeBERTa with sequential analysis using LSTM. In addition, the system supports image-based inputs through OCR, enabling detection of cyberbullying from chat screenshots and memes. The developed Gradio-based user interface allows users to input text or upload images and receive real-time predictions along with confidence scores. The interface first extracts text from the input (direct text or OCR from images), processes it through the trained model, and displays whether the content is toxic or non-toxic along with the corresponding probability. This helps in identifying harmful content quickly and efficiently in real-world scenarios. Figure 3 shows an example where normal conversational text extracted from an image is classified as non-toxic. The system correctly identifies that the conversation does not contain any harmful or abusive content, demonstrating its ability to handle general user interactions.

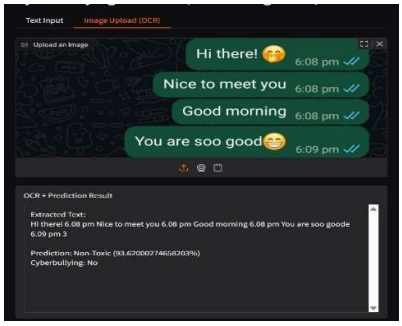

Figure 3: Prediction by the model as non-toxic in Gradio (Image OCR Input)

Figure 4 presents a case where the input contains threatening and abusive language. The model successfully classifies it as toxic, indicating the presence of cyberbullying. This demonstrates the system's ability to detect explicit harmful content.

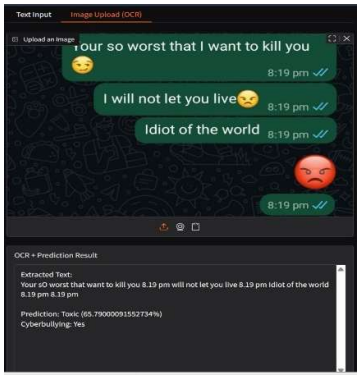

Figure 4. Prediction of the model as Toxic in Gradio (Image OCR Input)

Figure 5 presents an example containing highly offensive and abusive language. Compared to the toxic instance shown in Figure 4, the model assigns a substantially higher toxicity confidence score, indicating its ability to distinguish varying levels of harmful content severity. This demonstrates that the proposed framework not only identifies toxic content but also responds appropriately to stronger abusive expressions.

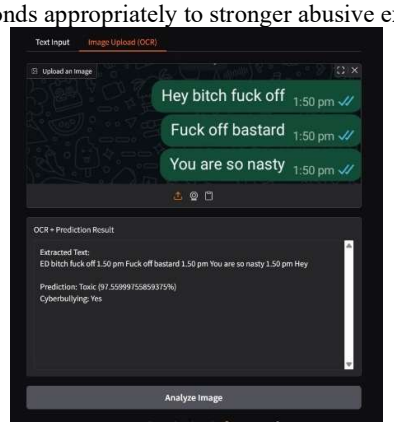

Figure 5. High Confidence Toxic Prediction

V. CONCLUSION

This paper introduces a hybrid multimodal method for detecting cyberbullying by integrating contextual comprehension through DeBERTa with sequential analysis via LSTM. The proposed system attained a validation accuracy of 94%, showcasing its ability to differentiate between toxic and non-toxic content. The

model demonstrates strong performance in identifying harmful content by understanding both contextual significance and sequential trends in text [8]. In addition to textual analysis, the system includes Optical Character Recognition (OCR) to retrieve text from images and accommodates emoji and visual feature interpretation, facilitating the identification of cyberbullying from chat screenshots and multimedia sources. A Gradio-based interface that is easy to use has been created to enable real-time classification, permitting users to enter text or upload images and receive instant results along with confidence scores. In general, the suggested approach strikes a solid equilibrium between precision and practical effectiveness, rendering it appropriate for real-world use. The combination of contextual, sequential, and multimodal analysis enables the system to identify both explicit and implicit types of cyberbullying effectively. In the future, the system can be improved by integrating sophisticated multimodal fusion methods, broadening to multi-class classification, and boosting performance with larger and more varied datasets.

REFERENCES

[1] R. Karim, S. Rahman, and M. Hasan, "Multimodal Hate Speech Detection from Bengali Memes and Texts," arXiv preprint, 2022, doi:10.48550/arXiv.2204.10196

[2] B. Murshed, M. Islam, and A. Rahman, "Deep Learning Based Multimodal Cyberbullying Detection System," Multimedia Tools and Applications, 2023

[3] P. Yi, L. Zhang, and H. Chen, "Session-Based Cyberbullying Detection in Social Media," Array, vol. 17, 2023, doi:10.1016/j.array.2023.100209

[4] Md. T. Ahmed, N. Akter, M. Rahman, A. Z. M. T. Islam, and D. Das, "Multimodal Cyberbullying Meme Detection from Social Media Using Deep Learning Approach," International Journal of Computer Science and Information Technology, vol. 15, no. 4, pp. 27–40, 2023, doi:10.5121/ijcsit.2023.15403

[5] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, and D. Testuggine, "The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes," Advances in Neural Information Processing Systems, 2020, doi:10.48550/arXiv.2005.04790

[6] K. Maity, S. Saha, and P. Bhattacharyya, "A Multi-task Multimodal Framework for Cyberbullying Detection in Memes," in Proc. SIGIR Conference, 2022, doi:10.1145/3477495.3531815

[7] P. Jha, A. Kumar, and R. Kumar, "Memeingful Analysis: Multimodal Explanation for Cyberbullying Detection," in Proc. EACL, 2024, doi:10.48550/arXiv.2401.09899

[8] N. M. Singh, R. Kumar, and S. Sharma, "Multi-modal Cyberbullying Detection with Severity Analysis," International Journal of Computer Network and Information Security, vol. 17, no. 3, pp. 65–75, 2025, doi:10.5815/ijcnis.2025.03.08

[9] G. M. Mohiuddin, M. Rahman, and S. Islam, "Deep Learning Models for Multimodal Cyberbullying Detection Using OCR," Multimedia Tools and Applications, 2025, doi:10.1007/s44163-025-00577-2

[10] A. Roy, S. Das, and P. Dutta, "Multi-class Cyberbullying Classification on Image and Text Data," Data Analytics Journal, 2025, doi:10.1016/j.dajour.2025.100067

[11] F. Barbieri, J. Camacho-Collados, L. Espinosa-Anke, and S. Schockaert, "Multimodal Emoji Prediction," in Proc. ACL, 2018, doi:10.48550/arXiv.1803.02392