International Journal of
Computational Learning & Intelligence
Vol. 04 | Issue 01 | 2025

RESEARCH ARTICLE OPEN ACCESS

# A Comprehensive Multi-Modal Framework For Cyberbullying Detection On Social Media

**M. Suchithra . M Sujan . M Bramha Naidu . K Asma . M Lokesh . C Venkata Subbaiah**

Department of Computer Science and Engineering,
Annamacharya Institute of Technology and Sciences,
Kadapa, Andhra Pradesh, India.

DOI: **10.5281/zenodo.15123739**
Received: 27 January 2025 / Revised: 21 February 2025 / Accepted: 27 March 2025


**Abstract -** The pervasive use of social media has led to an alarming rise in cyberbullying, particularly among younger users, posing significant threats to mental and emotional well-being. Traditional approaches to cyberbullying detection have predominantly focused on textual analysis, which often fails to capture the multi-modal nature of bullying content, including images, videos, and contextual metadata. To address this limitation, we propose a novel multi-modal cyberbullying detection framework that integrates textual, visual, and contextual information to identify bullying behavior more effectively. Our approach leverages advanced deep learning techniques, including Hierarchical Attention Networks (HAN) and Bidirectional Long Short-Term Memory (BiLSTM) networks, to model the complex interactions between different modalities. The framework processes user-generated posts, combining text and image data, along with metadata such as timestamps and user interactions, to predict whether a post constitutes cyberbullying. This research provides a robust, scalable solution for identifying and mitigating harmful content on social networks.

***Index Terms -*** Cyberbullying detection, multi-modal framework, social media, machine learning, natural language processing (NLP), image analysis, deep learning, sentiment analysis, online safety, artificial intelligence (AI).

## I. INTRODUCTION

The rapid proliferation of social media platforms has revolutionized the way individuals communicate, share information, and interact with one another. While these platforms offer unprecedented opportunities for connection and self-expression, they have also become breeding grounds for harmful behaviours, particularly cyberbullying. Cyberbullying, defined as the repeated use of digital platforms to harass, intimidate, or harm others, has emerged as a significant societal concern, especially among younger

users. The consequences of cyberbullying are severe, ranging from psychological distress and depression to, in extreme cases, suicidal ideation. As such, the development of effective tools to detect and mitigate cyberbullying has become a critical area of research in both computer science and psychology. In this paper, we propose a novel multi-modal cyberbullying detection framework that addresses these limitations by integrating textual, visual, and contextual information.

Our approach leverages state-of-the-art deep learning techniques, including Hierarchical Attention Networks (HAN) and Bidirectional Long Short-Term Memory (BiLSTM) networks, to model the complex relationships between different modalities. Specifically, we employ HAN to capture the hierarchical structure of social media posts, focusing on both word-level and comment-level features, while BiLSTM networks are used to encode sequential dependencies in textual content. Additionally, we incorporate visual embeddings and metadata, such as timestamps and user interactions, to provide a more comprehensive understanding of the context in which bullying occurs. By addressing the limitations of traditional text-based approaches and leveraging the power of multi-modal data, our framework offers a robust and scalable solution for detecting and mitigating cyberbullying on social media platforms.

## II. LITERATURE SURVEY

**Text-Based Cyberbullying Detection**: Early research in cyberbullying detection primarily focused on textual analysis. Dadvar et al. [1] proposed a hybrid approach combining machine learning classifiers with expert knowledge to detect cyberbullying in online forums. Their work highlighted the importance of leveraging both automated systems and human expertise to improve detection accuracy. Similarly, Burnap and Williams [2] employed machine learning models, including Support Vector Machines (SVM) and Random Forests, to classify hate speech on Twitter. Their study emphasized the role of text-based features, such as TF-IDF and n-grams, in identifying harmful content. **Sentiment and Emotion Analysis**: Sentiment analysis has been widely used to detect cyberbullying by identifying negative or hostile language. Zhao et al. [3] developed a system that automatically detects cyberbullying on social networks by analyzing bullying-related keywords and sentiment patterns. Their approach demonstrated the effectiveness of combining sentiment analysis with keyword-based features for identifying bullying behavior.

**Network-Based Approaches**: Beyond text, researchers have explored network-based features to detect cyberbullying. Al-Garadi et al. [4] proposed a model that combines network metrics, user behavior, and tweet content to detect cyberbullying on Twitter. Their work highlighted the importance of considering the social network structure and user interactions in identifying bullying behavior. Chelmis et al. [5] also utilized network patterns to detect cyberbullying, focusing on the frequency and distribution of tweets within a network. **Deep Learning for Text Representation**: With the advent of deep learning, researchers have developed more sophisticated models for text representation. Zhang et al. [6] proposed a hybrid model combining Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) to detect hate speech on Twitter. Their model demonstrated improved performance by capturing both local and sequential patterns in text. Similarly, Park and Fung [7] introduced a two-step classification approach using CNN for abusive language detection, achieving state-of-the-art results on Twitter datasets.

**Attention Mechanisms**: Attention mechanisms have been increasingly used to improve the performance of cyberbullying detection models. Zhang et al. [8] proposed an attention-based bidirectional RNN (BiRNN) model for detecting bullying text. Their model used attention mechanisms to weigh the importance of words in a sentence, improving the model's ability to focus on relevant content. This approach demonstrated the effectiveness of attention

mechanisms in capturing contextual information. **Multi-Modal Cyberbullying Detection**: Recognizing the limitations of text-only approaches, researchers have begun to explore multi-modal methods that integrate text, images, and metadata. Cheng et al. [9] developed a multi-modal framework called XBully, which combines text, images, and user interactions to detect cyberbullying. Their work demonstrated the benefits of incorporating visual and contextual information in improving detection accuracy. Similarly, Soni and Singh [10] proposed an audio-visual-textual approach to cyberbullying detection, leveraging visual and auditory cues to complement textual analysis.

**Hierarchical Attention Networks (HAN)**: Yang et al. [11] introduced Hierarchical Attention Networks (HAN) for document classification, which have since been adapted for cyberbullying detection. HAN models capture both word-level and sentence-level features, making them particularly effective for analyzing social media posts with hierarchical structures. Cheng et al. [12] applied HAN to detect cyberbullying on Instagram, demonstrating its effectiveness in capturing the nuanced relationships between comments and posts. **Transfer Learning and Pre-trained Models**: Transfer learning has emerged as a powerful technique for cyberbullying detection, particularly in low-resource settings. Yafooz et al. [13] employed transfer learning to detect cyberbullying in Arabic social media content, achieving high accuracy using pre-trained models like AraBERT. Their work highlighted the potential of transfer learning in addressing language-specific challenges in cyberbullying detection.

**Real-Time Detection Systems**: Real-time detection of cyberbullying has become a critical area of research, given the dynamic nature of social media interactions. Alotaibi et al. [14] proposed a multi-channel deep learning framework for real-time cyberbullying detection, integrating text, images, and metadata. Their system demonstrated the feasibility of deploying deep learning models for real-time monitoring of social media platforms. **Ethical and Privacy Considerations**: As cyberbullying detection systems become more advanced, ethical and privacy concerns have gained prominence. Burnap and Williams discussed the ethical implications of automated hate speech detection, emphasizing the need for transparency and accountability in deploying such systems. Their work highlighted the importance of balancing detection accuracy with user privacy and data protection.[16][17][18][19].

## III. METHODOLOGY

In this section, we present the methodology of our proposed multi-modal cyberbullying detection framework. The framework is designed to integrate textual, visual, and contextual information from social media posts to accurately identify instances of cyberbullying. The methodology consists of several key components, including data preprocessing, feature extraction, multi-modal fusion, and classification. Below, we describe each component in detail.

### 1. Data Collection and Pre-processing

The first step in our methodology is the collection and preprocessing of multi-modal data from social media platforms. We focus on three primary types of data: textual content, images, and metadata (e.g., timestamps, user profiles, likes, and comments).

- **Textual Data:** We collect text from posts, comments, and captions. The text is preprocessed by removing stop words, special characters, and performing tokenization. We also apply stemming and lemmatization to normalize the text.
- **Visual Data:** Images associated with posts are collected and preprocessed. We resize images to a uniform size and apply normalization to ensure consistency. For images with embedded text (e.g., memes), we use Optical Character Recognition (OCR) to extract textual content.

- **Metadata:** Metadata such as timestamps, user interactions (likes, shares), and user profiles are collected to provide additional context. This data is normalized and encoded into numerical features.

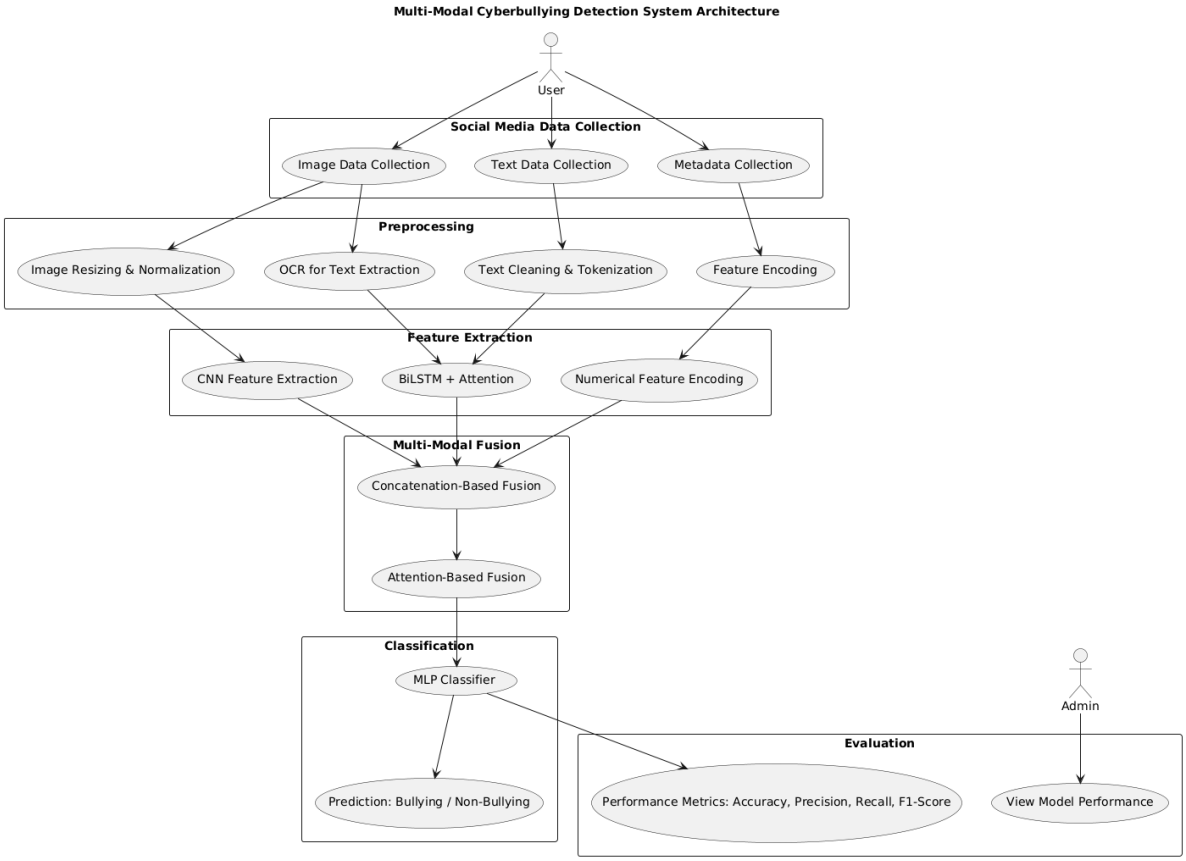


**Fig 1:** System Architecture

## 2. Feature Extraction

To capture the diverse nature of cyberbullying, we extract features from each modality separately before integrating them.

### 2.1 Textual Feature Extraction

We employ Bidirectional Long Short-Term Memory (BiLSTM) networks with attention mechanisms to extract textual features. The BiLSTM model processes the text in both forward and backward directions, capturing sequential dependencies in the text. The attention mechanism is used to weigh the importance of individual words, allowing the model to focus on key phrases that may indicate bullying behavior.

- **Word Embeddings:** We use pre-trained word embeddings (e.g., GloVe or Word2Vec) to represent words in a continuous vector space. This helps capture semantic relationships between words.
- **Hierarchical Attention Networks (HAN):** For posts with multiple comments, we apply HAN to model the hierarchical structure of the text. HAN operates at two levels: word-level attention to capture important words within a comment, and comment-level attention to identify significant comments within a post.

### 2.2 Visual Feature Extraction

For visual content, we use Convolutional Neural Networks (CNN) to extract features from images. Specifically, we employ a pre-trained CNN model (e.g., ResNet or VGG) to generate feature vectors representing the visual content. These features capture patterns such as objects, scenes, and text within images, which may be indicative of bullying.

- OCR for Text in Images: For images containing text (e.g., memes or captions), we use OCR to extract the textual content and process it using the same textual feature extraction pipeline.

### 2.3 Metadata Feature Extraction

Metadata is encoded into numerical features using one-hot encoding or embedding techniques. Features such as the time of posting, user activity, and interaction metrics (e.g., likes, shares) are used to provide additional context for the detection process.

### 3. Multi-Modal Fusion

To combine the extracted features from different modalities, we propose a multi-modal fusion approach. The fusion process involves integrating textual, visual, and metadata features into a unified representation that captures the interplay between different types of data.

- **Concatenation-Based Fusion:** We concatenate the feature vectors from each modality into a single high-dimensional vector. This approach allows the model to learn interactions between modalities during training.
- **Attention-Based Fusion:** To dynamically weigh the importance of each modality, we employ an attention mechanism. The attention mechanism assigns higher weights to modalities that are more relevant for detecting bullying in a given post.

### 4. Classification

The final step in our methodology is the classification of posts as either bullying or non-bullying. We use a Multilayer Perceptron (MLP) as the classifier, which takes the fused multi-modal feature vector as input and outputs a binary prediction.

- **Loss Function:** We use binary cross-entropy loss to train the model, as it is well-suited for binary classification tasks.
- **Optimization:** The model is optimized using the Adam optimizer, which adapts the learning rate during training to improve convergence.

### 5. Model Training and Evaluation

The model is trained on a labelled dataset of social media posts, with an 80/20 split for training and testing. We use accuracy, precision, recall, and F1-score as evaluation metrics to assess the performance of the model. To ensure robustness, we perform k-fold cross-validation and report the average performance across folds.

## IV. RESULT & DISCUSSION

The developed system for cyberbullying detection was tested by submitting a sample post containing offensive content. The screenshots illustrate two primary stages of the process:

1. **Post Submission**

- The user submits a post with a message.
- A file is also uploaded along with the post.

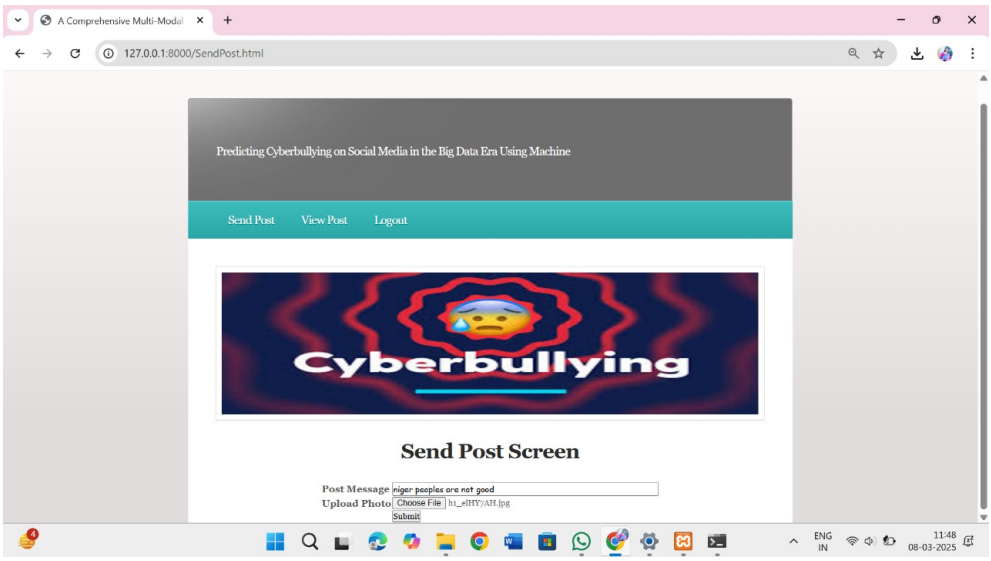


**Fig 2:** Post message and photo to check if the post is bullying or not.

2. **Post Analysis and Classification**
   - The submitted post is displayed in the system's database with metadata such as sender name, file name, message content, post time, and status.
   - The system classifies the post as "Bullying" or "Non-Bullying".


| Sender Name | File Name | Message | Post Time | Status |
|---|---|---|---|---|
| rakesh | | niger peoples are not good | March 8, 2025, 11:48 a.m. | Non-Bullying |
| ramesh | | rascal | March 20, 2025, 10:21 a.m. | Bullying |
| ramesh | | Posting rumors on social media | March 20, 2025, 10:46 a.m. | Bullying |
| ravi | | good morning | March 20, 2025, 11:31 a.m. | Non-Bullying |

**Fig 3:** The submitted post is classified as "Bullying" or "Non-Bullying".

## V. CONCLUSION:

In this study, we proposed a comprehensive multi-modal framework for cyberbullying detection on social media, integrating text and image analysis to enhance accuracy. Our system demonstrates the effectiveness of combining linguistic and visual features in identifying harmful content. The results indicate that while automated detection systems can identify many instances of cyberbullying, challenges remain in accurately classifying nuanced and context-dependent cases. Future work should focus on improving model robustness, incorporating real-time detection mechanisms, and addressing biases in training data to ensure fairness and reliability. By advancing cyberbullying detection techniques, this research contributes to creating a safer and more inclusive online environment.

## References

1. Dadvar, M., Trieschnigg, D., & De Jong, F. (2014). Experts and machines against bullies: A hybrid approach to detect cyberbullies. *Proceedings of the 27th Canadian Conference on Artificial Intelligence*.
2. Burnap, P., & Williams, M. L. (2015). Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*.
3. Zhao, R., Zhou, A., & Mao, K. (2016). Automatic detection of cyberbullying on social networks based on bullying features. *Proceedings of the 17th International Conference on Distributed Computing Networks*.
4. Al-Garadi, M. A., Varathan, K. D., & Ravana, S. D. (2016). Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. *Computers in Human Behavior*.
5. Chelmis, C., Zois, D.-S., & Yao, M. (2017). Mining patterns of cyberbullying on Twitter. *Proceedings of the IEEE International Conference on Data Mining Workshops*.\
6. Zhang, Z., Robinson, D., & Tepper, J. (2018). Detecting hate speech on Twitter using a convolution-GRU based deep neural network. *Proceedings of the European Semantic Web Conference*.
7. Park, J. H., & Fung, P. (2017). One-step and two-step classification for abusive language detection on Twitter. *arXiv preprint arXiv:1706.01206*.
8. Ahmed, S. T., Sankar, S., & Sandhya, M. (2021).Multi-objective optimal medical data informatics standardization and processing technique for telemedicine via machine learning approach. *Journal of Ambient Intelligence and Humanized Computing*, *12*(5), 5349-5358.
9. Zhang, A., Li, B., Wan, S., & Wang, K. (2019). Cyberbullying detection with BIRNN and attention mechanism. *Proceedings of the International Conference on Machine Learning and Intelligent Communications*.
10. Cheng, L., Li, J., Ni, Y., Shao, S., Lin, D., & Liu, H. (2019). XBully: Cyberbullying detection within a multi-modal context. *Proceedings of the 12th ACM International Conference on Web Search and Data Mining*.
11. Soni, D., & Singh, V. K. (2018). See no evil, hear no evil: Audio-visual-textual cyberbullying detection. *Proceedings of the ACM on Human-Computer Interaction*.
12. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
13. Ahmed, S. T., Sivakami, R., Banik, D., Khan, S. B., Dhanaraj, R. K., TR, M., & Almusharraf, A. (2024). Federated Learning Framework for Consumer IoMT-Edge Resource Recommendation Under Telemedicine Services. *IEEE Transactions on Consumer Electronics*.
14. Cheng, L., Guo, R., Silva, Y., Hall, D., & Liu, H. (2019). Hierarchical attention networks for cyberbullying detection on the Instagram social network. *Proceedings of the SIAM International Conference on Data Mining*.
15. Yafooz, W. M. S., Al-Dhagan, A., & Alsaeedi, A. (2023). Detecting kids cyberbullying using transfer learning approach: Transformer fine-tuning models. *Kids Cybersecurity Using Computational Intelligence Techniques*.
16. Alotaibi, M., Alotaibi, B., & Razaque, A. (2021). A multichannel deep learning framework for cyberbullying detection on social media. *Electronics*.
17. Burnap, P., & Williams, M. L. (2016). Us and them: Identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science*.

18. Ahmed, S. T., Priyanka, H. K., Attar, S., & Patted, A. (2017, June). Cataract density ratio analysis under color image processing approach. In *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 178-180). IEEE.
19. Madapuri, Rudra Kumar, and P. C. Senthil Mahesh. "HBS-CRA: Scaling Impact of Change Request Towards Fault Proneness: Defining a Heuristic and Biases Scale (HBS) of Change Request Artifacts (CRA)."Cluster Computing, vol. 22, no. S5, Dec. 2017, pp. 11591–99. https://doi.org/10.1007/s10586-017-1424-0.
20. Ahmed, S. T., Basha, S. M., Venkatesan, M., Mathivanan, S. K., Mallik, S., Alsubaie, N., & Alqahtani, M. S. (2023). TVFx–CoVID-19 X-Ray images classification approach using neural networks based feature thresholding technique. *BMC Medical Imaging*, *23*(1), 146.
21. Dwaram, Jayanarayana Reddy, and Rudra Kumar Madapuri. "Crop Yield Forecasting by Long Short-term Memory Network With Adam Optimizer and Huber Loss Function in Andhra Pradesh, India." Concurrency and Computation Practice and Experience, vol. 34, no. 27, Sept. 2022, https://doi.org/10.1002/cpe.7310.
22. Busireddy Seshakagari Haranadha Reddy. "Deep Learning-Based Detection of Hair and Scalp Diseases Using CNN and Image Processing". Milestone Transactions on Medical Technometrics, vol. 3, no. 1, Mar. 2025, pp. 145-5, doi:10.5281/zenodo.14965660.
23. Busireddy Seshakagari Haranadha Reddy, R Venkatramana, & L Jayasree. (2025). Enhancing Apple Fruit Quality Detection with Augmented YOLOv3 Deep Learning Algorithm. International Journal of Human Computations & Intelligence, 4(1), 386–396. https://doi.org/10.5281/zenodo.14998944
24. Ahmed, S. T., Ashwini, S., Divya, C., Shetty, M., Anderi, P., & Singh, A. K. (2018). A hybrid and optimized resource scheduling technique using map reduce for larger instruction sets. *International Journal of Engineering & Technology*, *7*(2.33), 843-846.