International Journal for Modern Trends in Science and Technology
Volume 12, Issue 04, pages 149-154.
ISSN: 2455-3778 online
Available online at: http://www.ijmtst.com/vol12issue04.html
DOI: https://doi.org/10.5281/zenodo.19324538

Design and Implementation of AI-Powered Cyberbullying Detection System

M Vijaya Bhaskar, Shaik Kamil Ahmed, Purandla Hari Pujitha, Maddasani Ramu, Gumperla Suvarshitha, Nelapati Sravanthi

Department of CSE-Artificial Intelligence, PBR Visvodaya Institute of Technology and Science, Kavali, Andhra Pradesh, India.

To Cite this Article
M Vijaya Bhaskar, Shaik Kamil Ahmed, Purandla Hari Pujitha, Maddasani Ramu, Gumperla Suvarshitha & Nelapati Sravanthi (2026). Design and Implementation of AI-Powered Cyberbullying Detection System. International Journal for Modern Trends in Science and Technology, 12(04), 149-154. https://doi.org/10.5281/zenodo.19324538

Article Info
Received: 28 February 2026; Revised: 18 March 2026; Accepted: 22 March 2026.

KEYWORDS

Cyberbullying, Multimodal Learning, Natural Language Processing, Convolutional Neural Networks, Real-time Monitoring, API Integration.

ABSTRACT

As social media platforms continue to grow, cyberbullying has become a major challenge affecting user safety and mental well-being. Detecting harmful content in real time is therefore essential. This paper presents a machine learning–based cyberbullying detection system that analyzes both textual and visual data. The proposed approach uses a multimodal framework combining Natural Language Processing (NLP) for text analysis and Convolutional Neural Networks (CNNs) for image classification. By integrating these techniques, the system effectively identifies abusive language, threatening messages, and inappropriate images with improved accuracy.

The system also enables real-time content moderation through API integration with external platforms such as LinkedIn and WhatsApp. Before content is posted, it is sent to the detection system for validation. If the content is harmful, it is blocked and necessary actions such as user flagging or profile restriction are applied. This proactive approach prevents cyberbullying at the source and enhances online safety.

INTRODUCTION

Cyberbullying has evolved from simple text-based abuse into a complex issue involving images, coded language, and cross-platform interactions. With the increasing use of visual content, there is a need for systems that can analyze both text and images simultaneously to understand the context of online behavior. Existing moderation techniques are largely reactive, where harmful content is removed only after it has been published, causing potential damage to users. To address this limitation, this paper proposes a proactive AI-based cyberbullying detection system.

The proposed system combines Natural Language Processing (NLP) and Convolutional Neural Networks (CNNs) to develop a multimodal classifier capable of identifying both textual and visual forms of

cyberbullying. In addition, a wireless API-based verification mechanism is introduced, which acts as a security layer between the user and social media platforms. Content is analyzed before being published, shifting the approach from post-detection to pre-prevention. The system also incorporates user profiling to enabling effective control and accountability across integrated applications.

OBJECTIVE

The main objective of this project is to develop an intelligent system capable of detecting and preventing cyberbullying in real time. The system integrates with external platforms such as LinkedIn, WhatsApp, and similar applications through secure APIs. Before any content is posted, it is transmitted to the detection system via a network for analysis. If the content is identified as safe, it is allowed to be published; otherwise, it is blocked, and actions such as user flagging or profile restriction are applied. This proactive approach helps prevent harmful content at the source rather than detecting it after publication. Overall, the system aims to enhance online safety, reduce harmful interactions, and create a secure digital environment using machine learning and real-time verification techniques.

LITERATURE SURVEY

Cyberbullying detection has become an important research area due to the rapid growth of social media platforms and the increasing amount of user-generated content. Early approaches focused on traditional machine learning algorithms such as Naïve Bayes, Support Vector Machines (SVM), and Decision Trees. These methods relied on feature extraction techniques like Bag-of-Words and TF- IDF to convert text into numerical form. Although these models achieved moderate accuracy, they often failed to understand context, sarcasm, and implicit abusive language, limiting their effectiveness.

To address these limitations, researchers introduced Natural Language Processing (NLP) and deep learning techniques. NLP methods involve preprocessing steps such as tokenization, stemming, and stop-word removal, along with sentiment analysis to detect emotional tone. Deep learning models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer-based models like BERT have shown improved performance by capturing contextual and semantic relationships in text. These approaches provide better accuracy and can detect complex patterns of cyberbullying.

Recent studies have also explored multimodal approaches that combine both text and image analysis to improve detection accuracy. Techniques such as Optical Character Recognition (OCR) and CNNs are used to identify harmful content in images, memes, and screenshots. Despite these advancements, many existing systems still follow a reactive approach, detecting cyberbullying only after content is posted. Therefore, there is a need for proactive systems that can analyze and filter content in real time. The proposed system addresses this gap by integrating multimodal learning with API-based real-time verification to prevent cyberbullying before it occurs.

NEED OF STUDY

The rapid growth of social media platforms has led to a significant increase in cyberbullying, posing serious threats to users' mental health and well-being. Cyberbullying includes abusive language, threats, and inappropriate content that can have long-term psychological effects. Existing systems rely mainly on manual moderation or post-detection techniques, which are inefficient for handling the large volume of real-time data generated on modern platforms. As a result, harmful content often reaches users before any corrective action is taken, highlighting the need for automated and proactive solutions.

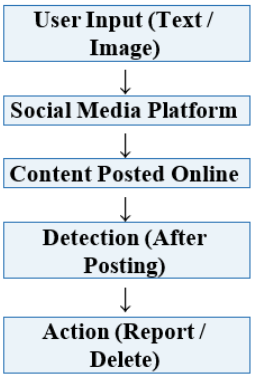
This study focuses on developing an AI-based cyberbullying detection system that uses a multimodal approach to analyze both text and images for improved accuracy. By integrating machine learning and deep learning techniques, the system can identify complex patterns of abusive behavior. Furthermore, the use of API-based integration allows content to be verified before posting, ensuring that harmful messages are blocked in advance and necessary actions are taken against users. This approach enhances online safety, reduces the spread of cyberbullying, and promotes a more secure digital environment.

EXISTING SYSTEM

Existing cyberbullying detection systems primarily rely on manual moderation and basic automated filtering techniques. Social media platforms typically use keyword- based filtering, rule-based approaches, or user reporting mechanisms to identify harmful content. However, these methods analyze content only after it has been posted, allowing abusive messages, harmful language, or inappropriate images to reach users before any action is taken. Additionally, most traditional systems focus mainly on textual data and do not effectively handle visual content such as images and videos, which are increasingly used in cyberbullying.

Furthermore, existing systems lack real-time prevention capabilities and are not integrated with external platforms for pre-verification of content. Their reliance on simple keyword matching reduces accuracy, as they fail to capture context, sarcasm, or hidden intent, leading to false positives and false negatives. As a result, these systems follow a reactive approach, making them less effective in preventing cyberbullying.

Existing System Workflow:



Disadvantages:

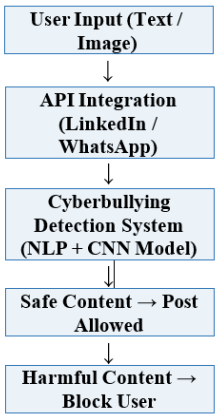
- No real-time content verification
- Detects harmful content only after posting
- Relies heavily on manual moderation
- Limited to text-based detection (no image analysis)
- Low accuracy due to keyword-based filtering
- Cannot detect context or hidden abuse
- No integration with external applications
- Ineffective in preventing cyberbullying at the source

PROPOSED SYSTEM

The proposed system introduces an AI-based cyberbullying detection mechanism that utilizes both machine learning and deep learning techniques to analyze textual and visual data. It follows a structured pipeline involving preprocessing, feature extraction, model training, and evaluation. Natural Language Processing (NLP) is used for text analysis, while Convolutional Neural Networks (CNNs) are used for image classification. This multimodal approach enhances detection accuracy by capturing both linguistic and visual patterns of cyberbullying.

A key feature of the system is real-time content verification through API integration with external applications such as LinkedIn, WhatsApp, and similar platforms. Before content is posted, it is transmitted to the detection system via a network. The system analyzes the content and determines whether it is safe or harmful. Safe content is allowed to be published, while harmful content is blocked, and appropriate actions such as warning or restricting the user are applied. This proactive approach prevents cyberbullying at the source. The system also includes user and service provider modules for effective monitoring and management of user activities.

Proposed System Workflow:



Advantages:

- Real-time cyberbullying detection
- Prevents harmful content before posting
- Uses AI (ML + DL) for high accuracy
- Multimodal analysis (text + images)
- Detects context and hidden abusive patterns
- API integration with platforms (LinkedIn, WhatsApp, etc.)
- Automatic blocking of harmful messages
- Ability to identify and restrict abusive users
- Scalable and efficient system
- Improves overall online safety

SYSTEM ARCHITECTURE

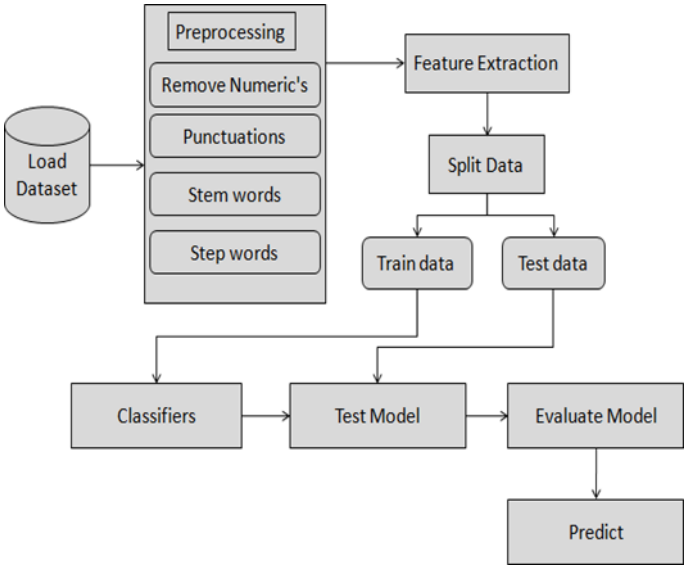


Fig. 1: System Architecture Diagram

SYSTEM MODULES

A. Service Provider Module

The Service Provider Module acts as the administrator of the system and manages the overall workflow. It includes functionalities such as login, dataset training and testing, viewing accuracy results, and monitoring user activities. This module ensures that the machine learning models are properly trained and updated to maintain high detection accuracy and system reliability.

B. Remote User Module

The Remote User Module allows users to interact with the system by registering, logging in, and submitting text or image content. Before posting, the content is analyzed to determine whether it is safe or abusive. The system also tracks user behavior and identifies repeated violations, enabling actions such as flagging or restricting user accounts to maintain a safe environment.

C. Database Module

The Database Module is responsible for storing all system data, including user details, datasets, model information, and analysis results. It provides secure storage and efficient data retrieval, ensuring smooth system operation. It also maintains records of abusive users for monitoring and further analysis.

D. API Integration Module

The API Integration Module enables communication between the detection system and external platforms such as LinkedIn and WhatsApp. Before content is posted, it is sent to the detection system via API for real-time verification. Based on the response, the content is either allowed or blocked, and necessary actions such as user restriction are applied. This module plays a key role in proactive cyberbullying prevention by filtering harmful content before publication.

SYSTEM REQUIREMENTS

Hardware Requirements:

The system requires moderate computational resources for data processing, model training, and real-time prediction. The hardware requirements are:

- Processor – Intel Core i3 or higher
- RAM – 8 GB minimum
- Hard Disk – 500 GB storage for datasets, models, and application files
- Graphics Processing Unit (Optional) – For faster deep learning and CNN processing
- Internet Connectivity – Required for API communication and integration with external applications

Software Requirements:

The system is developed using tools that support machine learning and web integration. The requirements are:

- Operating System – Windows 10 or higher (Linux/MacOS can also be used)
- Programming Language – Python (for machine learning and backend logic)
- Frontend Technologies – HTML, CSS, React (for user interface design)
- Backend Technologies – Python (Flask/Django /Spring boot framework recommended)
- Database – MySQL (for storing user data, results, and logs)
- Development Tools – Tomcat (for local server and database management)
- IDE/Editor – VS Code / PyCharm / Jupyter Notebook

TECHNIQUES USED IN THE PROJECT

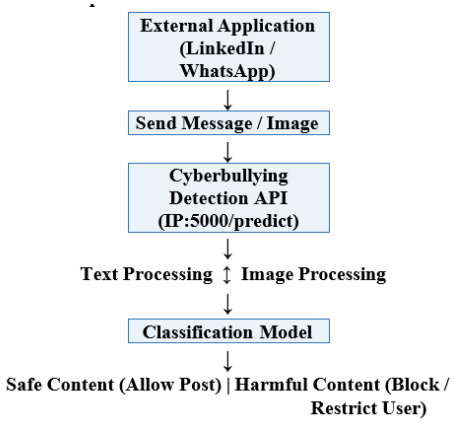
The proposed system utilizes a combination of machine learning, deep learning, and distributed system integration techniques to detect and prevent cyberbullying in real time. A key approach is the use of multimodal learning, which integrates both textual and visual data analysis. For text, Natural Language Processing (NLP) techniques such as tokenization, stop-word removal, stemming, and TF-IDF feature

extraction are applied to convert raw text into numerical form. Machine learning classifiers are then used to identify abusive or harmful language. For image analysis, Convolutional Neural Networks (CNNs) are employed to detect inappropriate or offensive visual content. This multimodal approach improves detection accuracy compared to single-modality systems.

The system also uses supervised machine learning techniques, where the dataset is preprocessed and divided into training and testing sets. Classification algorithms such as Logistic Regression, Support Vector Machines, or Neural Networks are used to classify content as safe or harmful. The performance of the model is evaluated using metrics such as accuracy, precision, recall, and F1-score. Once trained, the model is deployed as a prediction service that can analyze user input in real time, enabling automated content moderation without manual intervention.

To support real-time prevention, the system adopts a RESTful API-based architecture using frameworks such as Flask or FastAPI. The detection model is deployed on a local server (e.g., IP:5000/predict), which acts as a central verification unit. External applications such as LinkedIn or WhatsApp-like systems send user content to this API before posting. Based on the response, content is either allowed or blocked, and actions such as user restriction can be applied. The system operates over a local network (Wi-Fi), where devices communicate using the server's IP address, ensuring low latency and efficient data transfer. Security measures such as JSON-based communication, JWT authentication, and HTTPS can be used to ensure secure data exchange.

Detection Pipeline Flow:



RESULTS — USE CASE DIAGRAM

The system was validated through a use case diagram illustrating the interactions between three actors: the Service Provider, the Remote User, and the System. The Service Provider can register, log in, train and test the dataset, view accuracy results, identify abuser profiles, and manage remote users. The Remote User can register, log in, and submit content for verification. Key use cases include register, login, train and test dataset, view trained and tested accuracy results, identify the abuser profile, view all remote users, and logout.

CONCLUSION

In this work, an AI-powered cyberbullying detection system has been developed to enhance online safety by identifying and preventing harmful content in real time. The system adopts a multimodal approach by combining Natural Language Processing (NLP) for text analysis and Convolutional Neural Networks (CNNs) for image classification, enabling accurate detection of abusive language, threatening messages, and inappropriate images. Supervised machine learning models are used to learn cyberbullying patterns and provide reliable predictions.

A key contribution of the system is its integration with external platforms such as LinkedIn and WhatsApp through RESTful APIs. Before content is posted, it is transmitted to the detection system for analysis. Based on the result, content is either allowed or blocked, and actions such as user restriction can be applied. This proactive approach prevents harmful content at the source, improving user safety.

Overall, the proposed system provides an efficient and scalable solution for cyberbullying detection. It can be further enhanced by improving model accuracy, supporting multiple languages, and deploying the system in cloud environments for large-scale applications.

FUTURE ENHANCEMENT

Early studies focused on traditional machine learning algorithms such as Naïve Bayes, Support Vector Machines (SVM), Decision Trees, and Random Forests for detecting abusive language in textual data. These approaches relied on feature extraction techniques like Bag-of-Words and TF-IDF to convert text into numerical representations. Experimental results showed that SVM and deep learning models achieved higher accuracy, with deep learning methods reaching up to 96% accuracy in some cases. However, these methods often

lacked the ability to understand context, sarcasm, and implicit abusive language.

With advancements in technology, researchers began incorporating Natural Language Processing (NLP) techniques to improve detection accuracy. NLP-based models use preprocessing steps such as tokenization, stemming, and stop-word removal to analyze text effectively. Some studies have proposed sentiment analysis and modified TF-IDF techniques to better capture the emotional tone and intent of messages, improving classification performance.

Recent research has shifted towards deep learning approaches, which automatically learn features from large datasets. Models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer-based architectures like BERT have demonstrated superior performance compared to traditional methods. Deep learning models are capable of capturing complex patterns and contextual information, making them more effective for cyberbullying detection. Comparative studies also indicate that deep learning techniques significantly improve accuracy, precision, and recall in identifying harmful content.

In addition to text-based analysis, recent studies have explored multimodal approaches, combining text and image data for more comprehensive detection. Techniques such as Optical Character Recognition (OCR) and CNN- based image classification have been used to detect cyberbullying in images and screenshots. These methods help identify abusive content embedded in visual data, which is often overlooked in traditional systems.

Furthermore, researchers have proposed session-based and context-aware detection systems that analyze sequences of user interactions instead of individual messages. These approaches consider conversation context, user behavior, and temporal patterns to improve detection accuracy. Studies have also highlighted the importance of dataset quality, annotation techniques, and evaluation metrics in developing reliable models.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] N. Vidgen, A. Hale, S. Guest, H. Margetts, and D. Broniatowski, "Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection," Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 1667–1678, 2020.

[2] A. Risch and R. Krestel, "Aggression Identification Using Deep Learning and Data Augmentation," Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying, pp. 150–158, 2020.

[3] Z. Waseem, T. Davidson, D. Warmsley, and I. Weber, "Understanding Abuse: A Typology of Abusive Language Detection Subtasks," Proceedings of the Fourth Workshop on Online Abuse and Harms, pp. 1–13, 2021.

[4] A. Mathew, P. Saha, S. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, "HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 17, pp. 14867–14875, 2021.

[5] S. Mozafari, M. Farahbakhsh, and N. Crespi, "A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media," Complex Networks & Their Applications IX, pp. 928–940, 2020.

[6] J. K. Lee, H. S. Kim, and Y. J. Kim, "Cyberbullying Detection Using Deep Learning Models," IEEE Access, vol. 9, pp. 119519–119529, 2021.

[7] S. K. Saha, A. K. Pal, and M. S. Aktar, "Multimodal Cyberbullying Detection Using Deep Neural Networks," IEEE Access, vol. 10, pp. 88562–88574, 2022.

[8] M. R. Islam, M. M. Rahman, and M. R. Karim, "Cyberbullying Detection on Social Networks Using Machine Learning Techniques," IEEE Access, vol. 8, pp. 178372–178382, 2020.

[9] H. Al-Ahmad and M. Alsmadi, "Deep Learning Models for Cyberbullying Detection in Social Media," International Journal of Advanced Computer Science and Applications, vol. 12, no. 2, pp. 202–210, 2021.

[10] S. B. Khanday, M. A. Rabani, Q. Khan, and N. Rouf, "Deep Learning Based Approaches for Detecting Offensive Language in Social Media," IEEE Access, vol. 9, pp. 123453–123463, 2021.

[11] Y. Liu, M. Ott, N. Goyal et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv preprint arXiv:1907.11692, updated 2020.

[12] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," NAACL-HLT, updated applications widely used in 2020–2023.

[13] D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed., Draft, 2021.

[14] T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 38–45, 2020.

[15] S. Rajamanickam, P. M. Kumar, and V. Subramaniyaswamy, "Real-Time Cyberbullying Detection Using Deep Learning Models," IEEE International Conference on Artificial Intelligence and Knowledge Engineering, pp. 120–126, 2022.