International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0250
e ISSN: 2584-2854
Volume: 03
Issue:04 April 2025
Page No: 1540 - 1546

# Cyberbullying Detection and Prevention Using Machine Learning

*Mrs. Elakia K[1], Mr. Dinesh Kumar H[2], Mr. Daniyalraj K[3], Mr. Yogesh P[4], Mr. Vishva J[5]*
*[1]Professor, RGCET, Puducherry, India.*
*[2,3,4,5]B. Tech (CSE), RGCET, Puducherry, India.*
***Email ID:** elakia1692@gmail.com[1], kdaniyalaraj@gmail.com[3]*

## Abstract
*Cyberbullying Detection uses a combination of MACHINE LEARNING techniques such as TF-IDF vectorization, logistic regression, multilayer perception, CNNs and LSTM networks to create a robust model for detecting cyberbullying. By Employing BERT model, it's able achieve higher accuracy and better performance in identifying offensive content on social media platforms. The existing System for detecting cyberbullying in Indian Language Bengali on social media. The model uses text preprocessing, TF-IDF, and Instance Hardness Threshold (IHT) for resampling, it uses multiple Machine learning algorithms for detection of online harassment. However, the existing System does not address the practical challenges like Real-Time Detection and the Technique used for Resampling deduce the actual size of dataset to balance the dataset which leads to lower accuracy rate. To overcome these Limitations, the proposed system uses the BERT model, known for its advanced contextual understanding and bidirectional processing capabilities, to enhance prediction accuracy.*

***Keywords:** Linguistic Inquiry and Word Count, Cyberbullying, Psycholinguistic Tools, Deep Learning (DL): Techniques like CNNs, RNNs, and Bi-LSTM models.*

## 1. Introduction
Cyberbullying has become a significant concern in today's digital society, with widespread use of social media and messaging platforms amplifying its reach and impact. Unlike traditional bullying, cyberbullying allows individuals to target others through anonymous and continuous online interactions, which can lead to severe emotional and psychological distress. Victims of cyberbullying frequently experience anxiety, depression, and in extreme cases, may be driven to self-harm or suicide. Given the gravity of these consequences, timely and effective detection of cyberbullying has become essential. Detecting cyberbullying poses unique challenges due to the complexities of online language. Social media users often employ slang, sarcasm, and other colloquial expressions that can make harmful intent, making it difficult for traditional detection systems to capture context and identify abusive messages accurately. Existing methods, which typically rely on human moderators, struggle with the volume and speed of digital content, underscoring the need for automated systems capable of real-time analysis and response. To address these challenges, our research leverages the Bidirectional Encoder Representations from Transformers (BERT) model a powerful deep learning framework designed for nuanced language understanding [1][4]. BERT's bidirectional nature allows it to interpret context from surrounding words in a sentence, making it especially well-suited for detecting and context-dependent cyberbullying language. By training the model on a bilingual dataset of English and Tamil texts, we aim to broaden its applicability, allowing it to effectively identify cyberbullying across linguistic and cultural boundaries [2].

### 1.1 Challenges in Cyberbullying Detection
Identifying cyberbullying in digital content is challenging due to the nuanced nature of language, including slang, sarcasm, and cultural expressions. These factors complicate the identification process and often render traditional approaches ineffective [3]. Additionally, the sheer volume of online content requires systems that can process and analyze data in real-time. Existing models frequently struggle with accurately classifying nuanced language, leading to high false-positive and false-negative rates, which

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0250
e ISSN: 2584-2854
Volume: 03
Issue:04 April 2025
Page No: 1540 - 1546

undermine the effectiveness of automated detection [5].

### 1.2 Advances in Machine Learning for Natural Language Processing (NLP)

Recent advancements in NLP and Machine learning have made it possible to analyze language with greater context-sensitivity, which is crucial for effective cyberbullying detection. Models like the Bidirectional Encoder Representations from Transformers (BERT) represent a significant leap forward in language processing. BERT captures the context of words bidirectionally analyzing both preceding and succeeding words in a sentence—thereby improving its ability to interpret language nuances and subtleties. This development is essential for detecting implicit and indirect forms of cyberbullying.

### 1.3 Proposed System: BERT for Multilingual Cyberbullying Detection

Our study proposes a hybrid detection model leveraging BERT, focusing on both English and Tamil texts, to provide a culturally adaptive cyberbullying detection tool. BERT's bidirectional architecture enables a more comprehensive understanding of language, allowing for high accuracy in identifying harmful content. By including multilingual data, the model is equipped to detect various forms of bullying expressions across different linguistic and cultural contexts, addressing the limitations of previous models that were confined to a single language. Figure 1 shows Social Media Popularity Chart.

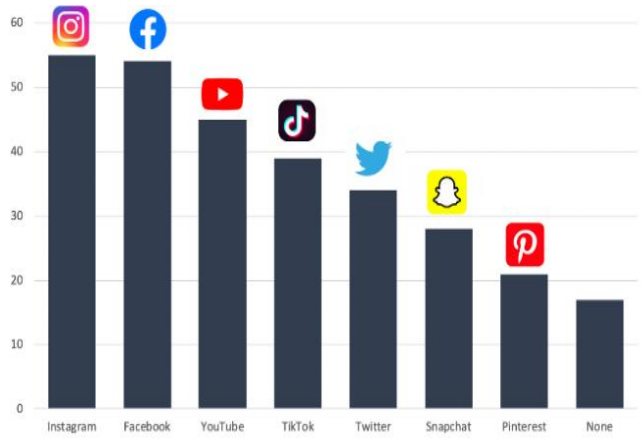


**Figure 1 Social Media Popularity Chart**

## 2. Literature Overview

### 2.1 Introduction to Cyberbullying Detection

Cyberbullying is a critical social issue amplified by the pervasive nature of digital platforms. The rise of social media has facilitated increased communication but also created new channels for harassment and bullying. Research has focused on understanding this issue, particularly given its significant psychological impact on victims [6]. As a response, computational approaches, such as machine learning and deep learning, have emerged to automatically detect harmful content, making real-time intervention possible. Early studies utilized traditional machine learning models such as Naive Bayes and Support Vector Machines (SVM), which showed promising results but were limited in handling the complexities of natural language used in online interactions.

### 2.2 Machine Learning Approaches

Machine learning models in cyberbullying detection rely on various linguistic and psychological features. Studies have shown that algorithms like Logistic Regression, Random Forest, and Decision Trees can identify patterns of abusive language effectively. Some research incorporated additional features, including sentiment analysis and emotion detection, to enhance accuracy [7]. Although traditional machine learning models achieved moderate success, their ability to fully capture the nuances of context in cyberbullying scenarios remains limited. This challenge led to exploring hybrid models that integrate conventional machine learning with advanced feature engineering and context-sensitive techniques, such as Instance Hardness Threshold (IHT) for resampling data.

### 2.3 Deep Learning Models and Their Advantages

Deep learning models like CNN, RNN, and more recently, BERT have been applied to cyberbullying detection with notable success. Unlike traditional machine learning methods, deep learning models automatically extract features from data, capturing complex patterns that are essential in identifying subtle forms of bullying, such as sarcasm or coded language. Studies highlight that CNN models are effective for image-based content classification, while RNNs excel in analyzing sequential data. The

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0250
e ISSN: 2584-2854
Volume: 03
Issue:04 April 2025
Page No: 1540 - 1546

integration of BERT, which processes text bidirectionally, has shown superior results in understanding context, making it highly suitable for detecting nuanced cases of cyberbullying [8].

### 2.4 Multilingual Detection in Cyberbullying Research

With the growing diversity of online communities, there is a need for multilingual detection systems. Research has extended beyond English to include languages such as English and Tamil Romanized dialects, addressing the unique cultural and linguistic features that influence bullying behaviors. In one study, a hybrid machine learning model tailored for English cyberbullying detection utilized the TfidfVectorizer for feature extraction and achieved a significant accuracy rate [9][10]. These multilingual approaches demonstrate the importance of creating adaptable detection models that cater to the linguistic diversity on digital platforms.

### 2.5 Challenges and Future Directions

Despite advancements, challenges remain in the consistent accuracy and ethical application of cyberbullying detection models. Studies report issues with data imbalance and the risk of false positives or negatives, which can undermine user trust and lead to inappropriate censorship. Further research is needed to address these limitations, with a focus on refining algorithms for context sensitivity and expanding datasets to improve model robustness. The development of hybrid models and transformer-based architectures like BERT, combined with multilingual capabilities, represents a promising direction for the next generation of cyberbullying detection tools.

## 3. Methodologies and Approaches

### 3.1 Multilingual Cyberbullying Detection

**Key Challenges:** Identifying offensive language across different languages is difficult due to linguistic nuances, variations in slang, and cultural differences.

**BERT-based Approaches:** BERT models can be trained on multilingual datasets to capture the nuances of different languages, enhancing the detection of context-dependent cyberbullying across diverse online communities.

**Application:** Multilingual models like BERT a show promise in handling multilingual text data, making them effective in social media.

### 3.2 Hybrid Machine Learning Models for Cyberbullying Detection

**Integration of Techniques:** Hybrid models combine traditional machine learning with deep learning, leveraging the strengths of both to improve detection accuracy.

**Feature Extraction:** These models utilize advanced feature extraction techniques like TF-IDF and word embeddings to capture significant linguistic features from text data, enhancing model predictions.

**Advantages:** Hybrid approaches help tackle class imbalance issues in datasets, improving the overall performance of cyberbullying detection systems.

### 3.3 Deep Learning Architectures in Cyberbullying Detection

**Convolutional Neural Networks:** CNNs are effective in extracting spatial features from text data, aiding in detecting abusive language in cyberbullying contexts.

**Recurrent Neural Networks (RNNs):** RNNs, particularly LSTMs and GRUs, excel in capturing sequential dependencies, making them suitable for analyzing time-series or conversational data in cyberbullying detection.

**BERT Models:** BERT's bidirectional training approach captures the full context of text, making it effective in understanding the subtleties of cyberbullying language, such as sarcasm or indirect threats.

### 3.4 Sentiment Analysis for Cyberbullying Detection

**Sentiment Analysis Integration:** By analyzing the sentiment of messages (positive, negative, or neutral), models can better differentiate between harmful and non-harmful content.

**Psycholinguistic Tools:** Tools like LIWC (Linguistic Inquiry and Word Count) and Empath's lexicon can be integrated to analyze emotional and psychological aspects of language in detecting cyberbullying.

**Improved Accuracy:** Incorporating sentiment analysis helps refine the model's ability to detect not just overtly abusive messages but also subtle, context-dependent forms of cyberbullying.

### 3.5 Challenges in Imbalanced Dataset Handling

**Class Imbalance:** Cyberbullying datasets often

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0250
e ISSN: 2584-2854
Volume: 03
Issue:04 April 2025
Page No: 1540 - 1546

suffer from class imbalance, where non-bullying content vastly outnumbers bullying instances, affecting model performance.

**Resampling Techniques:** Techniques like SMOTE (Synthetic Minority Over-sampling Technique) and Instance Hardness Threshold (IHT) help in balancing datasets by generating synthetic samples for the minority class or down-sampling the majority class.

**Performance Metrics:** Using precision, recall, F1-score, and ROC-AUC helps evaluate model performance effectively, especially in imbalanced datasets, to ensure a reliable detection system.

## 4. Findings and Trends

The cyberbullying detection using machine learning (ML) and deep learning approaches. Traditional ML techniques, like Support Vector Machines (SVM) and Logistic Regression, are moderately effective in identifying harmful language but struggle with nuanced cases due to their reliance on manually selected features. In contrast, deep learning models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) excel at automatic feature extraction and complex pattern recognition, allowing for more accurate detection of subtle abusive behaviors [11]. The BERT (Bidirectional Encoder Representations from Transformers) model further improves detection by processing text bidirectionally, capturing context-dependent meanings essential in cyberbullying cases. Notably, multilingual adaptation has become a key trend, with models now trained on diverse languages, such as English and Tamil which enhances their adaptability across cultures and language-specific expressions. Hybrid models that combine CNN and Bi-LSTM or attention mechanisms offer additional accuracy by leveraging both sequential and contextual data, which is useful in identifying time-dependent abusive language patterns. Table 1 shows Findings and Trends in cyberbullying detection and privation using ML.

### 4.1 Table

**Table 1 Findings and Trends in cyberbullying detection and privation using ML**

| Aspect | Details | Findings | Trends |
|---|---|---|---|
| Machine Learning vs. Deep Learning | - Machine Learning (ML): Traditional methods like SVM and Logistic Regression.<br>- Deep Learning (DL): Techniques like CNNs, RNNs, and Bi-LSTM models are utilized for better feature extraction. | Deep learning models are more effective in identifying complex, nuanced language patterns. | DL models are increasingly replacing ML for nuanced text classification tasks. |
| Role of BERT | BERT's bidirectional language processing is effective for understanding context. | BERT improves accuracy and sensitivity to context in cyberbullying detection. | BERT is widely adopted and increasingly applied across languages for a more robust understanding of context. |
| Hybrid Models | Combines CNN with Bi-LSTM or attention mechanisms to capture sequential and spatial data.<br>- Useful in detecting time-based or sequential patterns in language. | Hybrid models achieve higher accuracy, especially for abusive behavior that develops over multiple posts. | Combining multiple model types (e.g., CNN, LSTM) for enhanced robustness and versatility |
| Challenges | - Difficulty in interpreting sarcasm, slang, and rapidly evolving language.<br>- Computational demands for models like BERT in real-time applications.<br>- Privacy concerns and potential for false positives/negatives. | False positives and negatives remain a challenge, and computational costs can limit deployment. | Research into lightweight BERT variations (e.g., Distil BERT) and privacy-preserving approaches is growing. |

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0250
e ISSN: 2584-2854
Volume: 03
Issue:04 April 2025
Page No: 1540 - 1546

### 4.2 Architecture of BERT-Based Cyberbullying Detection System

- **Data Collection:** Collects raw text data from social media platforms.
- **Preprocessing Module:** Handles tokenization, lemmatization, and removal of noise like special characters.
- **Feature Extraction with BERT:** Processes text bidirectionally to capture the contextual meaning of words.
- **Classification Layer:** Utilizes algorithms (such as Logistic Regression or CNN) to classify content as bullying or non-bullying.
- **Output Layer:** Produces real-time results to flag potential cyberbullying content.

## 5. Challenges and Gaps

### 5.1 Contextual and Nuanced Language Understanding

Detecting sarcasm, slang, and context-dependent meanings poses a significant challenge. Cyberbullying language is often indirect, with abusive intent masked in subtle phrases or culturally specific slang, which is difficult for ML models to interpret accurately.

### 5.2 Computational Demands

Advanced models like BERT, while effective, are computationally intensive and require substantial resources. This poses difficulties for real-time deployment, especially on platforms with limited processing power or bandwidth.

### 5.3 False Positives and Negatives

ML models may sometimes misclassify benign content as bullying (false positives) or miss actual cases of bullying (false negatives). This challenge impacts user trust and system reliability, requiring improvements in model precision and recall.

### 5.4 Privacy Concerns

Cyberbullying detection requires monitoring user interactions, which can raise privacy issues. Balancing the need for surveillance with user privacy is a delicate issue that requires careful consideration and possibly the development of privacy-preserving models.

### 5.5 Adaptability Across Languages and Culture

Current models often struggle with multilingual and cross-cultural contexts, as certain words or phrases may have different connotations in different languages. Multilingual models are in development, but they still require extensive language-specific training to avoid bias and improve accuracy.

### 5.6 Class Imbalance in Datasets

Cyberbullying datasets often contain more non-bullying content than bullying instances, making it harder for models to detect minority classes. This imbalance can lead to lower recall for cyberbullying cases, reducing the system's effectiveness.

### 5.7 Dataset Limitations

Available datasets may not be comprehensive enough, especially in representing diverse languages, contexts, and evolving language patterns. This restricts the model's ability to generalize effectively to real-world cases, highlighting the need for more extensive and diverse datasets.

### 5.8 Difficulty in Real-time Detection

Real-time detection requires quick processing and minimal latency. Deep learning models that require significant processing time may struggle with real-time deployment, especially for high-traffic platforms.

### 5.9 Ethical Concerns with Automated Detection

Automating cyberbullying detection can lead to ethical dilemmas, such as censorship and potential biases. Automated systems must balance sensitivity with fairness, avoiding overly restrictive moderation that can limit freedom of expression.

### 5.10Rapid Evolution of Online Language

Slang and abusive language evolve quickly, often outpacing the adaptability of models. Models need frequent updates or adaptive mechanisms to keep up with changes in language and maintain high accuracy over time.

## 6. Chart

Representing the overall effectiveness and challenges of different aspects in cyberbullying detection approaches:

- **Deep Learning Accuracy:** High effectiveness due to automatic feature extraction. Figure 2 shows Cyberbullying Detection.
- **Machine Learning Accuracy:** Lower accuracy as it relies on manual feature selection.
- **Multilingual Effectiveness:** Shows strong adaptability in multilingual settings.

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0250
e ISSN: 2584-2854
Volume: 03
Issue:04 April 2025
Page No: 1540 - 1546

- **Psycholinguistic Sensitivity:** Moderate effectiveness in detecting emotional tones.
- **Hybrid Model Performance:** High performance by combining CNN, RNN, and attention mechanisms.
- **Challenges in Detection:** Moderate challenges due to issues like sarcasm, privacy, and resource demands.

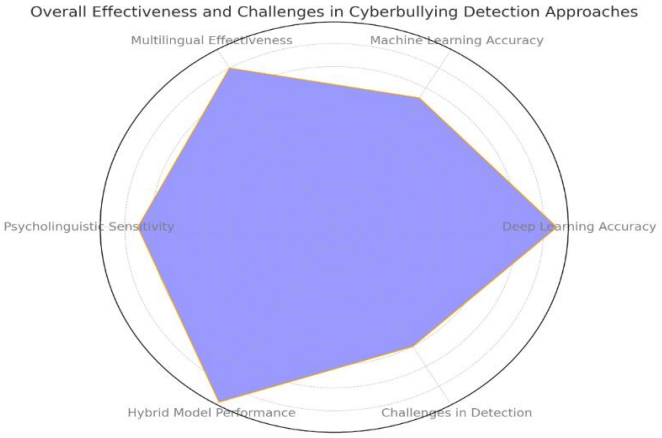


**Figure 2 Cyberbullying Detection**

## Conclusion

The current system effectively detects cyberbullying terms in Bengali using a deep learning model. We are Aiming to implement BERT model for better analysis outcome and also to implement Real-Time Detection while chatting by hosting a custom social media like chatting website(online) And also planning to implement Multi-Language detection of cyberbullying to maximize the efficiency of this project.

## References

[1]. Ahmed, N., Ahammed, R., Islam, M.M., Uddin, M.A., Akhter, A., Talukder, M.A., Paul, B.K., 2021d. Machine learning based diabetes prediction and development of smart web application. Int. J. Cogn. Comput. Eng. 2, 229–241.

[2]. Ahmed, M.F., Mahmud, Z., Biash, Z.T., Ryen, A.A.N., Hossain, A., Ashraf, F.B., 2021a. Bangla online comments dataset. Mendeley Data 1.

[3]. Ahmed, M.F., Mahmud, Z., Biash, Z.T., Ryen, A.A.N., Hossain, A., Ashraf, F.B., 2021b. Cyberbullying detection using deep neural network from social media comments in bangla language. arXiv preprint arXiv:2106.04506.

[4]. Ahmed, T., Mukta, S.F., Al Mahmud, T., Al Hasan, S., Hussain, M.G., 2022b. Bangla text emotion classification using LR, MNB and MLP with TF-IDF & CountVectorizer. In: 2022 26th International Computer Science and Engineering Conference. ICSEC, IEEE, pp. 275–280.

[5]. Ahmed, M.T., Rahman, M., Nur, S., Islam, A., Das, D., 2021c. Deployment of machine learning and deep learning algorithms in detecting cyberbullying in bangla and romanized bangla text: A comparative study. In: 2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies. ICAECT, IEEE, pp. 1–10.

[6]. Ahmed, M., Rahman, M., Nur, S., Islam, A., Das, D., et al., 2022a. Introduction of PMI SO integrated with predictive and lexicon based features to detect cyberbullying in bangla text using machine learning. In: Proceedings of 2nd International Conference on Artificial Intelligence: Advances and Applications. Springer, pp. 685–697.

[7]. Akhter, S., et al., 2018. Social media bullying detection using machine learning on bangla text. In: 2018 10th International Conference on Electrical and Computer Engineering. ICECE, IEEE, pp. 385–388.

[8]. Akter, M., Zohra, F.T., Das, A.K., 2017. Q-MAC: QoS and mobility aware optimal resource allocation for dynamic application offloading in mobile cloud computing. In: 2017 International Conference on Electrical, Computer and Communication Engineering. ECCE, IEEE, pp. 803–808.

[9]. Alkhatib, K., Abualigah, S., 2020. Predictive model for cutting customers migration from banks: Based on machine learning classification algorithms. In: 2020 11th International Conference on Information and Communication Systems. ICICS, IEEE, pp.

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0250
e ISSN: 2584-2854
Volume: 03
Issue:04 April 2025
Page No: 1540 - 1546

303–307.
[10]. Aurpa, T.T., Sadik, R., Ahmed, M.S., 2022. Abusive bangla comments detection on facebook using transformer-based deep learning models. Soc. Netw. Anal. Min. 12 (1), 1–14.
[11]. Bhattacharyya, S., Jha, S., Tharakunnel, K., Westland, J.C., 2011. Data mining for credit card fraud: A comparative study. Decis. Support Syst. 50 (3), 602–613.