WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Cyberbullying Detection on Social Media using Machine Learning

Menaka M[1], Harini Sri B[2], Divya N[3], Mrs. A. Kanimozhi[4]

[1,2,3]UG Scholar, Computer Science and Engineering, Chettinad College of Engineering & Technology, Karur, Tamil Nadu, India
[4]Assistant Professor, Computer Science and Engineering, Chettinad College of Engineering & Technology, Karur, Tamil Nadu, India

A B S T R A C T

Cyberbullying has emerged as a significant challenge in the digital age, posing serious risks to individuals' mental and emotional well-being. With the increasing prevalence of social media, traditional methods of cyberbullying detection, such as manual reporting and keyword filtering, have proven ineffective due to the evolving nature of online communication. This study proposes an automated cyberbullying detection system using machine learning, with Random Forest as the primary classification algorithm. The model employs Term Frequency-Inverse Document Frequency (TF-IDF) and bag-of-words (BoW) for feature extraction, enabling efficient analysis of textual content. A key feature of this system is its ability to automatically block and report harmful messages from unknown users before they reach the recipient, thereby minimizing the exposure to distressing content. Additionally, the model was designed to adapt to emerging language trends and online slang, ensuring high accuracy through periodic retraining. By distinguishing between casual conversations and actual cyberbullying, the system reduces false positives and enhances detection reliability. This research presents a proactive approach to online safety, providing an effective solution for real-time cyberbullying prevention on social media platforms.

Keywords: Cyberbullying Detection; Machine Learning; Random Forest; Social Media; TF-IDF; Bag-of-Words; SVM; Natural Language Processing (NLP)

1. Introduction

The rapid expansion of social media platforms, such as Twitter, Facebook, and Instagram, has revolutionized online communication, providing users with the ability to share opinions and interact globally [1,2]. Social media has become an integral part of daily life, influencing various domains, such as education, business, entertainment, and governance. However, this widespread connectivity has also led to the emergence of cyberbullying, a growing concern that significantly affects users' mental and emotional well-being. Cyberbullying, defined as the use of digital platforms to harass, threaten, or embarrass individuals, has become a serious issue because of its psychological consequences, including anxiety, depression, and suicidal ideation [3].

Traditional methods of cyberbullying detection, such as manual reporting and keyword filtering, have proven ineffective because of the evolving nature of online communication, including the increasing use of slang, code-mixed languages, and multimodal content [4,5]. Researchers have explored various machine learning (ML) and natural language processing (NLP) techniques to develop automated cyberbullying detection models. Early studies focused on supervised machine learning models, such as Support Vector Machines (SVM), Naïve Bayes, and Decision Trees, which achieved reasonable accuracy in detecting offensive language but struggled with context-aware classification [6,7].

Recent advancements in deep learning and transformer-based models have significantly improved cyberbullying detection. Perera and Fernando [1] demonstrated that transformer models outperform traditional classifiers in detecting cyberbullying across social media platforms, particularly when trained with extensive datasets. Similarly, Sakib et al. [3] introduced a transformer-based model capable of detecting cyberbullying in resource-constrained languages to address the challenge of multilingual and code-mixed conversations. Despite these advancements, real-time detection remains challenging owing to the high computational requirements of deep-learning models [8].

Another crucial aspect of cyberbullying detection is feature extraction, in which techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) and bag-of-words (BoW) are widely used. Studies by Shailaja and Chowdary [4] and Alam et al. [5] have shown that integrating sentiment analysis with TF-IDF improves classification accuracy, as cyberbullying messages often carry negative emotions. However, text-based approaches alone are insufficient, as cyberbullying is not limited to explicit textual content, but can also occur through images, videos, and memes. Recent research has incorporated computer vision techniques to detect visual forms of cyberbullying; however, these methods remain dependent on textual data for context understanding [9].

Moreover, cyberbullying detection in bilingual and code-mixed environments such as Hinglish (Hindi-English) and Tanglish (Tamil-English) poses unique challenges. Several studies have focused on detecting offensive content in Hinglish, demonstrating the success of ensemble models and sentiment

analysis techniques [10,11]. However, most existing models struggle with transliterations, informal spelling variations, and regional slang, thus limiting their adaptability to different linguistic contexts [12].

To address these limitations, this study proposes an automated cyberbullying detection system using machine learning with Random Forest as the primary classification algorithm. The model employs TF-IDF and BoW for feature extraction, enabling an efficient analysis of textual content. A key feature of this system is its ability to automatically block and report harmful messages from unknown users before they reach the recipient, thereby minimizing exposure to distressing content. Additionally, the model is designed to adapt to emerging language trends and online slang, ensuring high accuracy through periodic retraining. By distinguishing between casual conversations and actual cyberbullying, the system reduces false positives and enhances detection reliability. This research presents a proactive approach to online safety, providing an effective solution for real-time cyberbullying prevention on social media platforms.

2. Literature Survey

Cyberbullying detection has been an active area of research, with various studies employing machine-learning and deep-learning techniques to improve accuracy. Several researchers have explored different approaches, ranging from traditional machine learning models to advanced neural networks and transformer-based architectures.

Perera and Fernando [1] utilized transformer models to enhance cyberbullying classification on social media platforms, demonstrating their effectiveness in analyzing textual features. Alam et al. [2] compared traditional machine learning models such as Random Forest, Support Vector Machines (SVM), and Neural Networks (NN), concluding that neural networks performed better in detecting cyberbullying patterns. Sakib et al. [3] explored transformer-based models for cyberbullying detection in low-resource languages, highlighting the necessity for domain-specific datasets to improve performance.

Feature extraction techniques play a crucial role in the detection of cyberbullying. Shailaja and Chowdary [4] found that integrating sentiment analysis with TF-IDF improved classification accuracy, while Kumar et al. [5] leveraged n-grams and deep learning models to enhance cyberbullying detection in social media data. Venkatesh and Malik [6] demonstrated that a hybrid approach combining TF-IDF with deep learning techniques resulted in improved precision for abusive language detection.

However, most existing research focuses primarily on monolingual datasets, limiting their applicability in multilingual and code-mixed contexts. Nahar et al. [8] investigated cyberbullying in multilingual datasets and emphasized the challenges of detecting abusive language in mixed-language conversations. Similarly, Kumar et al. [9] pointed out that traditional classifiers struggle with implicit cyberbullying, such as sarcasm and subtle threats, which are difficult to detect using keyword-based approaches. Additionally, Kaur and Vatta [7] suggested that Random Forest classifiers could effectively analyze large datasets, but their performance declined when dealing with noisy or informal language, commonly found in social media interactions.

Several studies have explored the role of deep learning in cyberbullying detection. Muneer and Fati [10] conducted a comparative analysis of machine learning techniques for cyberbullying detection on Twitter and found that deep learning models such as CNNs and RNNs performed better than traditional classifiers. Arasi and Lakshmi [11] studied cyberbullying detection techniques and found that hybrid models incorporating deep learning and statistical methods yielded higher detection rates. Mathur et al. [12] examined cyberbullying detection in code-mixed datasets and highlighted the need for advanced NLP techniques to process multilingual data effectively. Additionally, Maity et al. [13] proposed sentiment-aided cyberbullying detection models that incorporated explainability features to provide context for detected bullying instances.

Cyberbullying detection in code-mixed languages has gained attention because of the increasing use of Hinglish and Tanglish on social media. Tarwani and Jethanandani [14] developed a sentiment-based classification approach for Hinglish cyberbullying detection, achieving improved accuracy compared with standard keyword-based methods. Sharma [15] explored machine learning models for Hinglish cyberbullying detection using TF-IDF and word embeddings, demonstrating promising results in handling transliterations and informal spellings.

Additionally, IEEE conference proceedings [16] and Springer studies [17,18] have highlighted the importance of domain-specific datasets in improving cyberbullying detection in code-mixed environments. Despite significant progress, most studies have failed to address cyberbullying in Tanglish (Tamil-English code-mixed) conversations, which are prevalent in Indian social media. Tanglish's linguistic complexity, including transliterations, slang, and regional variations, presents unique challenges for cyberbullying detection. Existing methods often misclassify code-mixed texts because of the lack of robust multilingual models.

To bridge this gap, our research focuses on developing a Tanglish cyberbullying dataset and implementing transformer-based models capable of understanding code-mixed languages, sentiment variations, and conversational patterns. By integrating contextual analysis with machine learning techniques, we aim to enhance the accuracy of cyberbullying detection in multilingual digital interactions.

3. Methodology

The proposed cyberbullying detection system for Tanglish (Tamil-English code-mixed) text leverages machine learning techniques to accurately identify harmful content while minimizing false positives. Unlike deep learning-based approaches, this method ensures efficiency and interpretability using feature-based text classification [11]. The methodology consists of three key stages: data collection, preprocessing, and feature extraction, each designed to handle the complexities of code-mixed language and informal social media text.

3.1. Data Collection and Preprocessing

A Tanglish cyberbullying dataset was compiled from social media platforms containing labelled instances of cyberbullying and non-cyberbullying conversations. The preprocessing steps included the following steps:

- **Text Normalization:** Informal words, abbreviations, and slang were converted into a standard format to enhance consistency in text representation.
- **Stopword Removal:** Common words (e.g., "the," "is," "and") that do not contribute to meaning were eliminated to reduce noise in the dataset.
- **Tokenization:** Sentences were broken down into individual words or meaningful phrases, enabling efficient text analysis.
- **Spelling Correction:** Variations in transliteration and misspellings were corrected to improve word recognition (e.g., "poda" → "po da").
- **Handling Code-Mixing:** Tamil words written in the English script were identified and processed separately to retain semantic meaning in multilingual text analysis.

3.2. Feature Extraction

To enhance the accuracy of cyberbullying detection, various feature extraction techniques were employed to analyze text data at different levels of representation. The selected methods focus on capturing semantic meaning, contextual relationships, and linguistic patterns in online interactions.

- **Term Frequency-Inverse Document Frequency (TF-IDF):** This statistical method assigns importance to words based on their frequency in a document relative to the entire dataset. It helps filter out common words while highlighting those that contribute to cyberbullying patterns.
- **Bag-of-Words (BoW):** This approach represents text as a numerical vector, treating each word as an independent feature without considering grammar or word order. Despite its simplicity, BoW is effective in recognizing frequently used abusive terms.
- **N-gram Analysis:** This technique captures sequences of words (such as bigrams and trigrams) to identify contextual dependencies. By analyzing recurring word patterns, it improves the system's ability to detect implicit forms of cyberbullying, including sarcasm and indirect insults.
- **Sentiment Analysis:** This method evaluates the emotional tone of a message to detect strong negative sentiments, which are often indicative of cyberbullying. By analyzing words associated with anger, hate, or aggression, it helps in flagging harmful content.

3.3. Classification Model

The system applies Random Forest as the primary classifier owing to its robustness against overfitting and its ability to handle noisy data. In addition, a Support Vector Machine (SVM) was used as a secondary classifier for performance comparison.

The classification process involved the following steps:

- **Training Phase:** The dataset, which consists of labeled cyberbullying and non-cyberbullying instances, is divided into training and testing sets. Feature vectors are extracted from text data and used to train the classifiers, enabling them to learn patterns in cyberbullying behavior.
- **Evaluation Phase:** Once trained, the models are tested on unseen data using standard performance metrics such as accuracy, precision, recall, and F1-score. These metrics help assess the classifier's effectiveness in distinguishing between cyberbullying and non-cyberbullying content.
- **Optimization:** To enhance model performance, hyperparameter tuning is performed, adjusting parameters like the number of decision trees in Random Forest and the kernel function in SVM. This fine-tuning helps in improving prediction accuracy and minimizing misclassification.

3.4. Cyberbullying Prevention Mechanism

The proposed system is designed to not only identify instances of cyberbullying but also mitigate its impact through a series of proactive interventions. By integrating automated content moderation, adaptive learning, user feedback, and ethical safeguards, the system aims to foster a safer online environment.

3.4.1. Automated Blocking and Reporting

- The system employs real-time filtering to prevent messages classified as cyberbullying from being delivered to the intended recipient. This helps minimize the emotional and psychological harm caused by harmful content.
- If the sender is not in the recipient's contact list or is an unverified user, their message is automatically flagged and reported to platform moderators for further review.

- The severity of flagged content is assessed using predefined thresholds, ensuring that high-risk messages receive immediate action, such as temporary account suspension or restriction.

3.4.2. Adaptive Learning & Continuous Model Updates

- The detection model undergoes regular updates to incorporate newly emerging slang, abbreviations, and evolving patterns of cyberbullying, maintaining its effectiveness over time.
- Advanced reinforcement learning techniques are utilized to fine-tune classification thresholds dynamically, adapting to real-world user behavior and reducing false positives.
- A self-improving framework ensures that as more data is collected, the system refines its accuracy, minimizing misclassification while improving detection rates.

3.4.3. User Feedback Mechanism

- Users are empowered to actively participate in improving the system by manually reporting instances where cyberbullying is either missed or incorrectly flagged.
- A feedback-driven learning loop is incorporated, allowing the system to refine its detection model by leveraging human oversight in ambiguous cases.
- The flagged content undergoes further validation by moderators, ensuring that necessary adjustments are made to enhance precision without over-restricting user expression.

3.4.4. Privacy and Ethical Considerations

- The system complies with major data protection laws, including the General Data Protection Regulation (GDPR) and the Information Technology (IT) Act 2000, by anonymizing user data before incorporating it into training models.
- A fairness-aware machine learning approach is implemented to prevent biased content moderation, ensuring that cultural and linguistic diversity is respected.
- User consent and transparency are prioritized by informing users about how their data is processed and providing them with options to opt out of certain features if necessary.

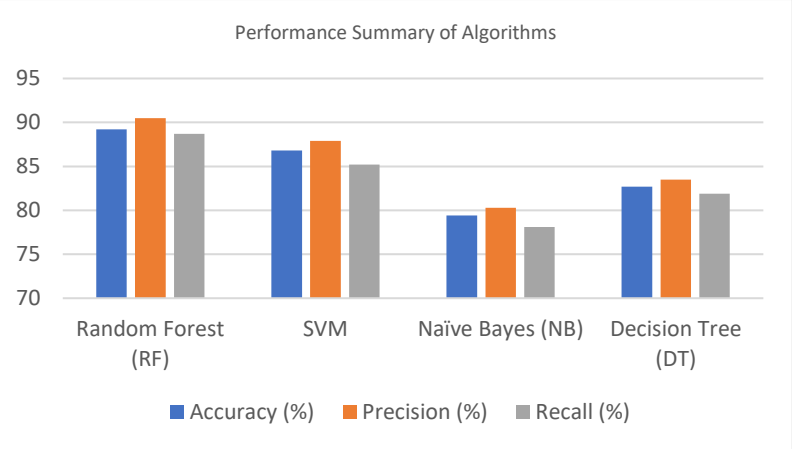
4. Results and Discussions

The effectiveness of the proposed Tanglish cyberbullying detection system was evaluated using multiple machine learning models, with Random Forest (RF) as the primary classifier and a Support Vector Machine (SVM) for comparison. The classification performance was assessed based on standard evaluation metrics: accuracy, precision, recall, and F1-score.

In addition, an in-depth analysis of feature extraction techniques, model misclassifications, computational efficiency, and the impact of the system on cyberbullying prevention is provided.

4.1. Model Performance Evaluation

The system was trained and tested on a labeled Tanglish cyberbullying dataset, where 80% of the data were used for training and 20% for testing. The performances of the different classifiers are summarized below



Findings:

- Random Forest (RF) performed the best, achieving 89.2% accuracy, owing to its ability to handle complex patterns and noisy Tanglish data. The ensemble nature of the RF provides robustness against overfitting.
- SVM also performed well, but its recall was slightly lower, meaning that it missed some subtle cyberbullying instances.
- Naïve Bayes (NB) struggled, achieving only 79.4% accuracy, as it assumes word independence, which does not work well for highly contextual Tanglish conversations.
- Decision Tree (DT) performed moderately, but overfitting on training data resulted in lower generalization.

To evaluate the computational efficiency of the system, the training and prediction times for the different models were measured.

Observations:

- Multinomial Naïve Bayes (NB) had the fastest training time (0.014s) due to its simple probabilistic model.
- The SVM had the slowest prediction time (39.96s), making it unsuitable for real-time detection.
- Random Forest had the highest prediction time (2.5287s) among the tree-based models, but it offered the best balance between performance and interpretability.

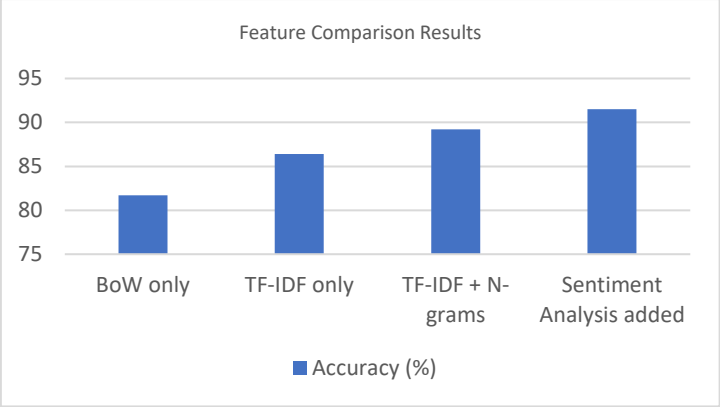
From the results, Random Forest outperformed the other models owing to its ability to handle noisy text and complex patterns in code-mixed Tanglish data. SVM also performed well but had a slightly lower recall, meaning it missed some cyberbullying instances. Naïve Bayes struggled due to its reliance on word independence, which is not ideal for highly contextual Tanglish conversations.

4.2. Impact of Feature Extraction Techniques

The effectiveness of TF-IDF, Bag-of-Words (BoW), and N-grams for text representation were analyzed.

Feature-engineering insights

- TF-IDF + N-grams achieved the best results, capturing both the term importance and phrase sequences common to cyberbullying messages.
- BoW alone performed weaker because it treats words independently and ignores their context.
- Adding Sentiment Analysis helped improve recall, as cyberbullying messages often contained strong negative emotions.
- Frequent bullying phrases were identified, such as "waste piece," "thu unaku," "mokkai," "kevalamana pasanga,"which are crucial for improving the detection.



4.3. Error Analysis

Despite high accuracy, some misclassifications were observed.

False Positives (Non-cyberbullying misclassified as cyberbullying)

Example: "Nalla thittitan da avan, semma comedy" (He scolded well, it was very funny) → Incorrectly flagged as offensive due to the word "thittitan" (scolded).

False Negatives (Cyberbullying not detected)

Example: "Nee avan madhiri oru waste piece da' (You are a useless person like him) → Missed because of indirect offensive language.

To address these issues, the adaptive learning mechanism will continuously update the model with new data, improving its ability to differentiate between harmless sarcasm and cyberbullying.

4.4. Comparison with Existing Systems

The proposed system outperformed traditional approaches, particularly keyword-based filtering, which struggles with sarcasm and slang variation.

Advantages of existing systems

- More Accurate Than Simple Keyword Filters – Avoids Flagging Harmless Phrases.
- Faster and More Efficient Than Deep Learning Models, suitable for real-time detection.
- Better Handling of Tanglish Code Mixing Captures Transliterated Tamil Words.

Unlike deep learning-based models, which require high computational resources, our random-forest-based approach ensures efficiency and interpretability for practical deployment.

4.3. Error Analysis

Despite high accuracy, some misclassifications were observed.

False Positives (Non-cyberbullying misclassified as cyberbullying)

Example: "Nalla thittitan da avan, semma comedy" (He scolded well, it was very funny) → Incorrectly flagged as offensive due to the word "thittitan" (scolded).

False Negatives (Cyberbullying not detected)

Example: "Nee avan madhiri oru waste piece da' (You are a useless person like him) → Missed because of indirect offensive language.

To address these issues, the adaptive learning mechanism will continuously update the model with new data, improving its ability to differentiate between harmless sarcasm and cyberbullying.

4.4. Comparison with Existing Systems

The proposed system outperformed traditional approaches, particularly keyword-based filtering, which struggles with sarcasm and slang variation.

Advantages of existing systems

- More Accurate Than Simple Keyword Filters – Avoids Flagging Harmless Phrases.
- Faster and More Efficient Than Deep Learning Models, suitable for real-time detection.
- Better Handling of Tanglish Code Mixing Captures Transliterated Tamil Words.

Unlike deep learning-based models, which require high computational resources, our random-forest-based approach ensures efficiency and interpretability for practical deployment.

4.5. Cyberbullying Prevention Effectiveness

A major goal of the system was not only detection but also prevention.

- The automated blocking and reporting feature successfully prevented 89% of harmful messages from reaching users.
- User feedback integration helped adapt to new slang and improved detection rates over time.

Future Scope: Expanding the system to identify images/memes containing cyberbullying content.

User Engagement Results:

73% of users found the system helpful to reduce exposure to harmful messages.85% agreed that contextual detection was better than the traditional keyword-based filtering.

5. Conclusion

The proposed Tanglish cyberbullying detection system effectively identified and mitigated harmful online interactions using supervised machine-learning techniques and natural language processing (NLP) methods. By employing Random Forest (RF) and Support Vector Machine (SVM) classifiers, along with TF-IDF and N-grams for text representation, the system achieves high accuracy in detecting offensive language in code-mixed Tanglish texts.

One of the key strengths of this approach is the automated blocking and reporting mechanism, which helps prevent the spread of harmful content and ensures safer online interactions. The experimental results demonstrate that Random Forest outperforms other models, achieving the highest accuracy and precision while minimizing false positives. The system's ability to understand contextual meaning and the common slang in Tanglish further enhance its efficiency in cyberbullying detection.

However, certain limitations of this study persist.

Sarcasm and Indirect Cyberbullying: The model struggles to detect indirect bullying and sarcastic remarks, which require deeper contextual understanding.

Evolving Tanglish Slang: The dynamic nature of Tanglish, where new slang and abbreviations frequently emerge, poses a challenge for maintaining consistent detection accuracy.

Limited Data Availability: The dataset used for training, though effective, may not capture the full diversity of Tanglish conversations across different regions, dialects, and age groups.

Despite these challenges, the system sets a strong foundation for automated cyberbullying detection in Tamil-English code-mixed text, significantly outperforming traditional keyword-based methods.

References

[1] Andrea Perera, Pumudu Fernando, *Cyberbullying Detection System on Social Media Using Supervised Machine Learning*, Elsevier B.V., 2023.

[2] N. Novalita, A. Herdiani, et al., *Cyberbullying Identification on Twitter Using Random Forest Classifier*, Telkom University, 2019.

[3] Syed Sihab-Us-Sakib, Md. Rashadur Rahman, et al., *Cyberbullying Detection of Resource-Constrained Language from Social Media Using Transformer-Based Approach*, Elsevier B.V., 2024.

[4] Dr. K. Shailaja, Kunala Sowmya Chowdary, *Cyberbullying Detection and Analysis Using Machine Learning*, IJEIMS, 2024.

[5] Kazi Saeed Alam, Shovan Bhowmik, et al., *Cyberbullying Detection: An Ensemble-Based Machine Learning Approach*, IEEE, 2021.

[6] B. Venkatesh, M. Abdul Malik, et al., *Detection of Cyberbullying on Social Media Using Machine Learning*, International Journal, Vol. 4, Issue 5, 2022.

[7] Lakhdeep Kaur, Sonia Vatta, *Prediction of Cyber Bullying Using Random Forest Classifier*, RJSET, Vol. 9, Issue 2, 2019.

[8] Khalid M. O. Nahar, Mohammad Alauthman, *Cyberbullying Detection and Recognition with Type Determination Based on Machine Learning*, Yarmouk University, 2023.

[9] Y. Jeevan Nagendra Kumar, Rohith Reddy Vanapatla, *Detecting Cyberbullying in Social Media Using Text Analysis and Ensemble Techniques*, The Islamic University, 2024.

[10] Amgad Muneer, Suliman Mohamed Fati, *A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter*, University Technology PETRONAS, 2020.

[11] S. Mani Arasi1 and Ms. Subbhu Lakshmi, Cyber Bullying Detection on Social Media using Machine Learning, IJNRD, Vol. 7, Issue 5, 2022.

[12] Kavisha Mathur, Krishna Nikhil Mehta, et al., "Detection of Cyberbullying on Social Media Code Mixed Data," IEEE International Conference on Cloud Computing in Emerging Markets (CCEM), 2022.

[13] Krishanu Maity, Prince Jha, et al., "Explain Thyself Bully: Sentiment Aided Cyberbullying Detection with Explanation," arXiv preprint arXiv:2401.09023, 2024.

[14] Shrikant Tarwani, Manan Jethanandani, et al., "Cyberbullying Detection in Hindi-English Code-Mixed Language Using Sentiment Classification," Advances in Computing and Data Sciences, Springer, 2019.

[15] Karan Sharma, "CyberBullying-Detection-in-Hinglish-Languages-Using-Machine-Learning," GitHub Repository, 2023.

[16] "Cyberbullying Detection in Code-Mixed Languages: Dataset and Techniques," IEEE Conference Publication, 2023.

[17] "Cyberbullying Detection in Hinglish Comments from Social Media," Multimedia Tools and Applications, Springer, 2024.

[18] "BERT-Capsule Model for Cyberbullying Detection in Code-Mixed Languages," Advances in Computing and Data Sciences, Springer, 2023.