International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 01 | Jan – 2026 DOI: 10.55041/ISJEM05359
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

# Cyberbullying Detection on Social Media Using Machine Learning

**V. MAGESWARI., MCA**
(Assistant Professor, Master of Computer Applications)
**G. THINAKARAN., MCA**
Christ College of Engineering and Technology
Moolakulam, Oulgaret Municipality, Puducherry – 605010.

**Abstract**

The rapid expansion of social media has transformed online communication, allowing users to share thoughts and opinions instantly. Along with these benefits, social platforms have also become a space where cyberbullying and abusive language spread quickly, causing serious emotional and psychological harm. Because social media content is large in volume, short in length, and highly informal, identifying harmful messages manually is both time-consuming and unreliable[1][2][6].This paper presents a machine learning–based system for detecting cyberbullying in social media text. The proposed approach processes user-generated content using Natural Language Processing techniques such as tokenization, stop-word removal, and text normalization[10][11]. The cleaned text is converted into numerical features using the TF–IDF method and classified using a Random Forest algorithm to determine whether the content is bullying or non-bullying[14][15]. A Flask-based web application is developed to provide real-time prediction through a simple and user-friendly interface[12]. The results show that the system can effectively identify harmful messages from short social media posts, making it a practical tool for improving online safety and supporting automated content moderation[1][5].

**Keywords**

Cyberbullying detection, social media analysis, machine learning, natural language processing, TF–IDF, Random Forest, text classification.

## 1. Introduction

The rapid growth of social media platforms and their widespread adoption have transformed the way people communicate and share information. While these platforms encourage connectivity and expression, they have also become major channels for cyberbullying, harassment, and abusive behavior. Cyberbullying includes offensive language, threats, humiliation, and targeted harassment, which can lead to serious psychological, emotional, and social consequences for victims. The informal nature of social media text, combined with its massive volume, makes effective monitoring and control a challenging task. Traditional content moderation methods, which rely heavily on manual review or keyword-based filtering, are often inefficient and fail to handle evolving language patterns, slang, sarcasm, and context-sensitive abuse[1][2][3].

These challenges have led to the increasing adoption of machine learning techniques for automated cyberbullying detection. Machine learning models are capable of learning complex patterns from large datasets and can identify harmful content beyond simple word matching. In text-based cyberbullying detection, Natural Language Processing plays a crucial role by transforming unstructured human language into structured numerical representations that can be processed by algorithms. Feature extraction techniques such as TF–IDF help capture the importance of words within short social media messages, enabling classifiers to distinguish between bullying and non-bullying content more effectively[4][6][10].

Building on this approach, the present work focuses on the design, implementation, and evaluation of a machine learning–based cyberbullying detection system. The proposed system applies systematic text preprocessing, TF–IDF feature extraction, and a Random Forest classification model to analyze social media messages. In addition, the system is implemented as a web-based application using the Flask framework to support real-time prediction and user interaction. This study aims to demonstrate that machine learning can provide an efficient and scalable solution for identifying cyberbullying content and improving safety in online social environments[5][12][14].

**The key contributions of this work are:**

1. Design and implementation of an automated cyberbullying detection pipeline that processes social media text and generates clear classification results with minimal user intervention[1][2].
2. Application and evaluation of a machine learning–based classification approach using TF–IDF feature extraction and a Random Forest model for detecting bullying and non-bullying content[5][14].
3. Detailed analysis of text preprocessing and feature representation techniques to handle short, informal, and noisy social media messages effectively[10][11][15].
4. Development of a web-based application that provides real-time cyberbullying prediction through a simple and user-friendly interface[12][13].
5. Modular system architecture implemented using Flask, allowing easy scalability, maintenance, and future integration with advanced models or platforms[12][14].

## 2. Materials and Methods

### 2.1 Dataset

This work leverages a publicly available **Kaggle cyberbullying dataset**, which provides labeled text samples representing bullying and non-bullying content collected from social media platforms[8]. The dataset contains short, informal user-generated messages suitable for Natural Language Processing and machine learning–based text classification. The dataset was divided into a training set of 80% and a testing set of 20% to ensure proper evaluation and to prevent model overfitting. After applying preprocessing steps such as text cleaning and feature filtering, the final test set contained a sufficient number of samples to support reliable performance

 

evaluation of the proposed cyberbullying detection system[1][6].

### 2.2 Static Feature Extraction

Feature extraction was performed using Natural Language Processing techniques to convert raw social media text into meaningful numerical representations suitable for machine learning[10][11]. Since social media messages are short, informal, and noisy, the extraction process focuses on capturing important linguistic patterns without altering the original meaning of the text. The analysis is non-intrusive and operates directly on user-generated text data. For each text instance, the following feature extraction steps were systematically applied:

- **Text Preprocessing:** The raw text was cleaned by converting it to lowercase, removing punctuation, special characters, and unnecessary symbols. Stop words were eliminated, and stemming was applied to reduce words to their root forms. This step helps reduce noise and standardize the input text[13][15].
- **Token-Based Features:** The cleaned text was tokenized into individual words, enabling the identification of meaningful terms that contribute to bullying or non-bullying behavior. Tokenization helps preserve word-level information critical for classification[13].
- **TF–IDF Vectorization:** The processed tokens were transformed into numerical feature vectors using the Term Frequency–Inverse Document Frequency (TF–IDF) technique. TF–IDF assigns weights to words based on their importance within a message and across the entire dataset, allowing the model to focus on discriminative terms while reducing the impact of commonly occurring words[10][15].

These features have been widely recognized in existing literature as effective for text classification and cyberbullying detection tasks. To evaluate the effectiveness of the extracted features, the machine learning model was trained and tested using the TF–IDF representations of the text data. This approach enables efficient handling of high-dimensional text features and provides reliable discrimination between bullying and non-bullying messages[14].

### 2.3 Machine Learning Models

A supervised machine learning algorithm was applied for the cyberbullying detection task using the extracted text features. The model was selected based on its effectiveness in text classification and its ability to handle high-dimensional feature spaces generated by TF–IDF vectorization[12][14].

**Random Forest Classifier:** Random Forest is an ensemble-based learning algorithm that builds multiple decision trees and combines their outputs to produce a final classification result. It is effective in handling noisy and unbalanced text data commonly found in social media platforms and helps reduce overfitting by aggregating predictions from multiple trees[14]. In this project, the Random Forest model was trained on TF–IDF feature vectors to classify text samples as bullying or non-bullying.

The model was trained using vectorized text features to perform binary classification, where bullying messages were labeled as one class and non-bullying messages as the other. Hyperparameter tuning was carried out to optimize model performance and improve prediction accuracy. Standard evaluation metrics such as accuracy, precision, recall, and F1-score were used to assess the effectiveness of the model and to analyze its classification behavior[5][12][15]. This machine learning–based approach demonstrates the suitability of supervised learning techniques for automated cyberbullying detection in social media environments[6][14].

### 2.4 System Architecture

The cyberbullying detection system is designed as a multi-tier web-based application to ensure modularity, scalability, and effective system management[12][14]. The architecture separates user interaction, processing logic, machine learning prediction, administration, and data storage into independent modules. The high-level system architecture consists of the following core components:

1. **Frontend Interface:** A web-based user interface developed using HTML and CSS that allows users to submit social media text for analysis. The interface displays the classification result indicating whether the content is bullying or non-bullying in a clear and user-friendly manner[12].
2. **Server:** The backend of the system is implemented using the Flask microframework. The Flask server handles client requests, manages application logic, and coordinates communication between the feature extraction, prediction, and database modules[12][13].
3. **Feature Extraction Engine:** This module processes the input text using Natural Language Processing techniques such as text cleaning, tokenization, stop-word removal, and stemming. The cleaned text is transformed into numerical feature vectors using the TF–IDF vectorization method, which serves as input to the machine learning model[10][11][15].
4. **Prediction Engine:** The prediction engine loads the pre-trained Random Forest model stored as a serialized .pkl file. It performs binary classification on the TF–IDF feature vectors to predict whether the given text contains cyberbullying content[14].
5. **Report Generator:** This module generates the prediction output and presents the result to the user through the web interface. It supports real-time analysis and provides a clear interpretation of the classification outcome.
6. **Database Layer:** SQLite is used as the database layer to store user inputs, prediction results, login credentials, and administrative records. The lightweight database ensures efficient data storage and easy maintenance[12].
7. **Administration Module:** The admin module provides authorized access for administrators to manage the system. It allows monitoring of user activity, viewing stored prediction records, and managing dataset or system-level operations. This module supports system maintenance and ensures controlled access to administrative functions[6][27].

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 01 | Jan – 2026 DOI: 10.55041/ISJEM05359
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

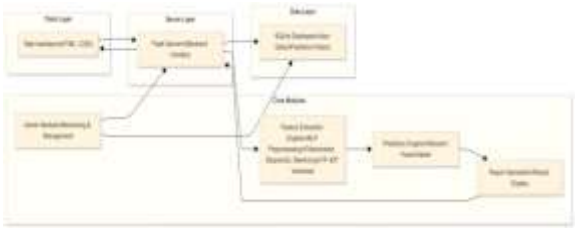

**Figure 1: Architecture of the Cyberbullying Detection System**

## 3. Results and Discussion

### 3.1 Feature Analysis and Importance

The feature extraction analysis revealed clear patterns in social media text that help differentiate cyberbullying content from non-bullying messages. Words related to insults, threats, harassment, and offensive expressions were found to appear more frequently in bullying samples than in normal conversations. Certain combinations of abusive terms and negative expressions were particularly indicative of cyberbullying behavior, while non-bullying text mainly contained neutral or positive language commonly used in daily communication[1][9][21]. The analysis of TF–IDF features showed that words associated with aggressive or harmful intent received higher importance scores due to their repeated occurrence in bullying messages and limited presence in non-bullying content. Short but emotionally strong phrases played a significant role in classification, as cyberbullying messages often rely on concise and direct language to convey harm. In contrast, non-bullying messages exhibited a broader vocabulary with less emotionally charged terms[10][11].Thus, the extracted text features provided strong discriminative information for identifying cyberbullying content. Proper text preprocessing combined with TF–IDF representation effectively captured meaningful linguistic patterns despite the noisy and informal nature of social media text. These findings reinforce the importance of resilient feature extraction techniques for handling evolving abusive language and improving the reliability of machine learning–based cyberbullying detection systems.

### 3.2 Model Performance Evaluation

Feature analysis plays a crucial role in understanding how textual patterns contribute to the identification of cyberbullying content in social media data. Since social media messages are short, informal, and often emotionally charged, effective feature representation is essential for accurate classification. In this work, text features were extracted using the TF–IDF technique after applying standard Natural Language Processing preprocessing steps[10][11]. The analysis revealed that words associated with insults, harassment, threats, and negative sentiment appeared more frequently in bullying samples than in non-bullying text. Such terms received higher TF–IDF weights because they occurred repeatedly in bullying messages while remaining relatively rare in normal conversations. In contrast, non-bullying content was dominated by neutral or positive words with lower discriminative power[1][9]. TF–IDF proved effective in highlighting important terms by balancing word frequency within individual messages and across the entire dataset. This representation reduced the influence of commonly occurring words and emphasized meaningful expressions related to abusive behavior. As a result, the Random Forest classifier was able to utilize these weighted features to distinguish bullying content more reliably[14][15].

**Table 1: Importance of Extracted Text Features**

| Feature Type | Description | Contribution to Detection |
|---|---|---|
| **Offensive Keywords** | Insults, abusive and threatening terms | High |
| **Negative Sentiment** | Words expressing anger, hate, or aggression | High |
| **Neutral Words** | Common conversational terms | Low |

The feature importance analysis confirms that text-based features derived from TF–IDF provide strong discriminative information for cyberbullying detection. Offensive and negatively charged words contribute significantly to classification decisions, while neutral and positive terms have minimal impact.

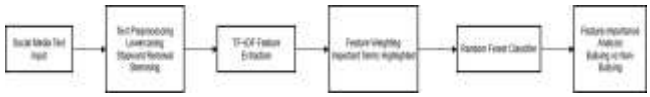

**Figure 2: Feature Extraction and Importance Flow**

**Figure Description:**

The diagram illustrates the feature extraction process used in the cyberbullying detection system. Social media text input undergoes preprocessing, followed by TF–IDF vectorization to generate weighted feature vectors. These features are then supplied to the Random Forest classifier, which assigns importance scores to different textual patterns during the classification process.

**Key findings:**

1. **Feature Effectiveness:**
TF–IDF–based text features proved to be highly effective in detecting cyberbullying content. Words related to insults, harassment, and negative sentiment contributed most to distinguishing bullying messages, while neutral and positive terms had minimal influence on classification decisions[1][9][10].
2. **Model Performance:**
The Random Forest classifier demonstrated reliable and consistent performance on high-dimensional TF–IDF feature vectors. Its ensemble-based nature allowed it to handle noisy and informal social media text effectively, resulting in balanced prediction behavior across bullying and non-bullying classes[14][15].
3. **Class-specific Insights:**
The precision–recall trade-off revealed the following characteristics:
   - Bullying-related keywords and negative expressions significantly improved the detection of harmful content, leading to strong recall for bullying messages[1][6].

 

- Neutral and positive language patterns helped reduce false positives, ensuring that non-bullying messages were not incorrectly flagged[4][11].
- Short but emotionally strong phrases played a critical role in identifying cyberbullying instances[3][9].

**Practical Implications:** The findings suggest that a text-driven approach using robust preprocessing and TF–IDF feature representation, combined with a Random Forest classifier, is well suited for real-world cyberbullying detection. This approach can be effectively deployed for automated content moderation to identify harmful messages while maintaining a balanced and reliable classification outcome[5][14][22].

### 3.3 System Implementation and Usability

The cyberbullying detection system was implemented using the Flask framework, which enabled rapid development and ensured a clear separation between the user interface, application logic, and machine learning components[12][13]. The lightweight nature of Flask allowed the system to efficiently handle user requests and perform real-time text analysis without noticeable delay, making it suitable for interactive use.The system processes user-submitted text through a structured pipeline that includes preprocessing, TF–IDF feature extraction, and prediction using the trained Random Forest model[10][14]. SQLite is used as the backend database to store user inputs, prediction results, and administrative records, providing a simple yet effective solution for data management[12]. The prediction results are generated instantly and displayed to users through the web interface.

Usability is enhanced through a clear and minimal output presentation. The system provides:

- A clear classification verdict indicating whether the input text is bullying or non-bullying.
- Immediate feedback based on machine learning prediction results.
- A structured workflow that allows users to interact with the system without technical complexity[1][5].

The admin module further improves system usability by enabling authorized users to monitor stored predictions, manage records, and oversee system activity. This administrative control supports system maintenance and ensures reliable operation. Overall, the implementation transforms the cyberbullying detection model into a practical, user-friendly web application that supports real-time analysis and effective content moderation[14][22].

### 3.4 Discussion of Limitations and Trade-offs

The primary advantage of the proposed cyberbullying detection approach lies in its simplicity, speed, and efficiency. Since the system relies on text-based analysis using TF–IDF features and a trained machine learning model, it is capable of processing social media messages in real time without requiring complex computational resources. This makes the approach suitable for large-scale content moderation, where a high volume of user-generated text must be analyzed quickly[6][10][12].

Despite these advantages, the system has certain inherent limitations.:

- **Context and Language Variability:** Cyberbullying language often changes over time and may include sarcasm, coded words, or context-dependent expressions that are difficult to detect using only textual features. As a result, some bullying messages may not be correctly identified[1][3][21].
- **Limited Semantic Understanding:** The model primarily relies on word frequency and weighting, which limits its ability to fully understand deeper semantic meaning or intent behind messages, especially in cases where harmful behavior is subtle[9][11][24].
- **Feature Dependence:** The current approach evaluates textual features independently, and interactions between words or phrases are not explicitly modeled. Combining multiple feature representations could potentially improve detection accuracy.

These limitations indicate that text-based machine learning models are best used as an initial filtering mechanism within a broader cyberbullying prevention strategy. While the current system provides reliable baseline performance, future enhancements involving contextual analysis and advanced models could further strengthen detection accuracy and robustness[5][14][30].

## 4. Conclusion

This study successfully designed, developed, and evaluated a machine learning–based cyberbullying detection system capable of identifying harmful content in social media text. The system utilizes Natural Language Processing techniques and TF–IDF feature extraction to represent textual data effectively, enabling accurate classification of bullying and non-bullying messages using a Random Forest classifier. The experimental results demonstrate that the proposed approach provides reliable performance when handling short, informal, and noisy social media content.

The implementation highlights the effectiveness of combining robust text preprocessing with an ensemble-based learning model to achieve balanced classification results. The Flask-based web application transforms the trained model into a practical and interactive system, offering real-time prediction and ease of use. The inclusion of an admin module further enhances system usability by allowing monitoring and management of stored predictions and system activity.

Overall, the results indicate that text-based feature representation plays a crucial role in cyberbullying detection, with offensive and negative expressions contributing significantly to classification decisions. The modular system architecture and lightweight implementation make the proposed solution suitable for deployment in real-world content moderation environments. This work provides a solid foundation for automated cyberbullying detection and supports the development of safer and more responsible online communication platforms.

## 5. Future Work

Future enhancements to overcome the current limitations and improve the effectiveness of the cyberbullying detection system include the following directions:

1. **Hybrid Feature Representation:**
Future work can explore the combination of multiple textual feature representations, such as TF–IDF with word embeddings or sentiment-based features, to capture richer linguistic and contextual information from social media text.
2. **Context-Aware Analysis:**
Incorporating contextual information such as conversation history, user interactions, and message sequences can improve the detection of subtle or implicit cyberbullying that may not be identifiable through isolated text analysis.
3. **Deep Learning Integration:**
Advanced deep learning models such as LSTM, CNN, or transformer-based architectures can be integrated to learn complex semantic relationships and long-range dependencies in text, potentially improving detection accuracy for evolving bullying patterns.
4. **Real-Time and Scalable Deployment:**
The system can be extended to support real-time deployment on cloud platforms using scalable architectures. This would allow the model to handle high volumes of social media data efficiently and support large-scale content moderation.
5. **Explainable AI Techniques:**
Integrating explainability methods such as LIME or SHAP can help interpret model predictions by highlighting influential words or phrases. This would increase transparency, build trust in automated decisions, and assist administrators in understanding detection outcomes.
6. **Robustness Against Evasion:**
Future research can focus on improving resistance to adversarial text manipulation, such as deliberate misspellings or coded language, by incorporating adaptive learning mechanisms and periodic model retraining.

## 6. References

[1] Dinakar, K., Reichart, R., & Lieberman, H. (2011). Modeling the Detection of Textual Cyberbullying. Proceedings of the International AAAI Conference on Web and Social Media (ICWSM).

[2] Dadvar, M., Trieschnigg, D., & de Jong, F. (2013). Improving Cyberbullying Detection with User Context. European Conference on Information Retrieval (ECIR).

[3] Xu, J. M., Jun, K. S., Zhu, X., & Bellmore, A. (2012). Learning from Bullying Traces in Social Media. Proceedings of NAACL-HLT.

[4] Nandhini, B. S., & Sheeba, J. I. (2015). Cyberbullying Detection and Classification Using Information Retrieval Algorithm. International Journal of Advanced Research in Computer and Communication Engineering.

[5] Agrawal, S., & Awekar, A. (2018). Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms. European Conference on Information Retrieval (ECIR).

[6] Salawu, S., He, Y., & Lumsden, J. (2017). Approaches to Automated Detection of Cyberbullying: A Survey. IEEE Transactions on Affective Computing.

[7] Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012). Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. International Conference on Privacy, Security, Risk and Trust.

[8] Kaggle. Cyberbullying Detection Dataset. Retrieved from https://www.kaggle.com

[9] Waseem, Z., & Hovy, D. (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. Proceedings of the NAACL Student Research Workshop.

[10] Schmidt, A., & Wiegand, M. (2017). A Survey on Hate Speech Detection Using Natural Language Processing. Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media.

[11] Fortuna, P., & Nunes, S. (2018). A Survey on Automatic Detection of Hate Speech in Text. ACM Computing Surveys.

[12] Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research.

[13] Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media.

[14] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.

[15] Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.

[16] Yin, D., Xue, Z., Hong, L., Davison, B. D., Kontostathis, A., & Edwards, L. (2009). Detection of Harassment on Web 2.0. Proceedings of the Content Analysis in the Web.

[17] Reynolds, K., Kontostathis, A., & Edwards, L. (2011). Using Machine Learning to Detect Cyberbullying. IEEE International Conference on Machine Learning and Applications.

[18] Van Hee, C., Lefever, E., & Hoste, V. (2015). Detection and Fine-Grained Classification of Cyberbullying Events. International Conference on RANLP.

[19] Hosseinmardi, H., Mattson, S. A., Rafiq, R. I., Han, R., Lv, Q., & Mishra, S. (2015). Detection of Cyberbullying Incidents on the Instagram Social Network. International Conference on Social Informatics.

[20] Rosa, H., Pereira, N., Ribeiro, R., Ferreira, P. C., Carvalho, J. P., Oliveira, S., Coheur, L., & Paulino, P. (2019). Automatic Cyberbullying Detection: A Systematic Review. Computers in Human Behavior.

[21] Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. Proceedings of ICWSM.

[22] Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep Learning for Hate Speech Detection in Tweets. Proceedings of the WWW Conference.

[23] Zhang, Z., Robinson, D., & Tepper, J. (2018). Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network. European Semantic Web Conference.

[24] Mozafari, M., Farahbakhsh, R., & Crespi, N. (2019). A BERT-based Transfer Learning Approach for Hate Speech Detection. Complex Networks.

[25] Al-Hassan, A., & Al-Dossari, H.. (2019). Detection of Hate Speech in Social Media Using Machine Learning. International Journal of Advanced Computer Science and Applications.

[26] Schmidt, A., & Wiegand, M. (2017). A Survey on Hate Speech Detection Using NLP. SocialNLP Workshop.

[27] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. KDD Conference.

[28] Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems (NeurIPS).

[29] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. ICLR Workshop.

[30] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT.