Automated Cyberbullying Detection Using Machine Learning With TF- IDF Features

Mr.Abdul Majeed
Assistant Professor
Vidya Jyothi Institute of Technology
Hyderabad
majeed@vjit.ac.in

G.Varshini
UG Student
Vidya Jyothi Institute of Technology
Hyderabad
varshinigajula0503@gmail.com

k.Shravya
UG Student
Vidya Jyothi Institute of Technology
Hyderabad
kshravya0202@gamil.com

S. Aishwarya
UG Student
Vidya Jyothi Institute of Technology
Hyderabad
aishwaryareddy1104@gmail.com

E.Sai Charan
UG Student
Vidya Jyothi Institute of Technology
Hyderabad
errollahari000@gmail.com

How to Cite this Article:

Charan, E., Aishwarya, S., k.Shravya, & G.Varshini, (2026). Automated Cyberbullying Detection Using Machine Learning With TF- IDF Features. International Journal of Creative and Open Research in Engineering and Management, 02(04). https://doi.org/10.55041/ijcope.v2i4.999

https://doi.org/ 10.55041/ijcope.v2i4.999

Abstract: Cyberbullying has become a serious social issue with the rapid growth of social media platforms. The massive volume of online user-generated content makes manual moderation ineffective and inconsistent. This research proposes an automated cyberbullying detection system using Machine Learning and Natural Language Processing techniques. Textual data is preprocessed and transformed using TF-IDF (Term Frequency–Inverse Document Frequency) vectorization to extract meaningful statistical features. Supervised classification algorithms such as Support Vector Machine (SVM) and Multinomial Naive Bayes are implemented to classify content as bullying or non-bullying.

Experimental evaluation demonstrates that TF-IDF combined with traditional machine learning classifiers provides an efficient and computationally lightweight solution for explicit abuse detection. The study also highlights the limitations of frequency-based approaches in handling sarcasm, contextual ambiguity, and evolving slang. The proposed model offers a scalable framework suitable for real-time deployment in social media platforms.

Keywords: Cyberbullying Detection, Machine Learning, TF-IDF, Natural Language Processing, Text Classification, SVM, Naive Bayes

1.Introduction:

The rapid growth of social media platforms and online communication tools has transformed the way individuals interact and share information. While these platforms provide numerous benefits, they have also become environments where harmful behaviours such as cyberbullying occur. Cyberbullying refers to the use of digital technologies to harass, threaten, insult, or humiliate individuals. Unlike traditional bullying, online harassment can spread quickly, reach a large audience, and remain accessible for long periods, causing significant psychological distress, anxiety, depression, and reduced self-esteem among victims.

The increasing volume of user-generated content on platforms such as messaging applications, forums, and social networking sites makes manual moderation highly impractical. Millions of comments, posts, and messages are generated every minute, making it impossible for human moderators to monitor and filter harmful content efficiently. Traditional keywordbased filtering systems are also insufficient, as users frequently employ slang, abbreviations, creative spellings, and sarcasm to bypass detection mechanisms. As a result, there is a critical need for intelligent automated systems capable of accurately identifying abusive and harmful language in real time.

Machine Learning (ML) and Natural Language Processing (NLP) offer promising solutions for addressing this challenge. By converting textual data into numerical representations using techniques such as Term Frequency–Inverse Document Frequency (TF-IDF), it becomes possible to analyze patterns in language statistically. Supervised learning algorithms can then classify content as bullying or non-bullying based on learned patterns from labelled datasets. This approach provides a scalable, efficient, and consistent method for detecting harmful online behaviour.

This research focuses on developing an automated cyberbullying detection system using TF-IDF feature extraction combined with traditional machine learning classifiers. The proposed framework aims to achieve high accuracy while maintaining computational efficiency, making it suitable for real-time deployment in social media environments. Through systematic experimentation and evaluation, the study demonstrates how statistical text representation can effectively contribute to safer digital communication platforms.

2. Literature Survey

Cyberbullying detection has become an important research area due to the rapid growth of social media platforms. Early research primarily focused on keyword-based filtering systems, where predefined lists of offensive or abusive words were used to detect harmful content. Although these systems were simple and easy to implement, they lacked contextual understanding and failed to detect disguised insults, sarcasm, and creative spellings. Researchers observed that keyword-based approaches produced high false positives and false negatives, making them unreliable for large-scale deployment.

To overcome these limitations, researchers introduced statistical feature extraction techniques such as Term Frequency–Inverse Document Frequency (TF-IDF). TF-IDF assigns weights to words based on their importance within a document relative to the entire dataset. Studies have shown that TF-IDF improves text representation by reducing the impact of common words while emphasizing unique and potentially harmful terms. This transformation of textual data into numerical vectors enabled the effective application of machine learning algorithms for classification tasks.

Several supervised machine learning algorithms have been evaluated for cyberbullying detection using TFIDF features. Support Vector Machines (SVM) and Logistic Regression have demonstrated strong performance in binary classification problems due to their ability to create optimal decision boundaries. Multinomial Naive Bayes has also been widely used because of its efficiency in handling word-frequencybased data. Comparative studies indicate that SVM often achieves higher accuracy, while Naive Bayes offers faster computation suitable for real-time applications.

In recent years, deep learning models such as Long Short-Term Memory (LSTM) networks and Transformer-based models like BERT have been introduced for cyberbullying detection. These models capture contextual and sequential relationships between words, enabling better understanding of sarcasm and implicit abuse. Research findings suggest that deep learning approaches generally outperform traditional machine learning methods in terms of contextual accuracy. However, they require large datasets, high computational power, and longer training time.

Despite the advancements in deep learning, traditional machine learning models combined with TF-IDF remain relevant due to their simplicity, interpretability, and lower computational cost. For

applications requiring real-time detection and limited hardware resources, TF-IDF with classifiers like SVM and Naive Bayes continues to provide an effective and practical solution. Therefore, this study builds upon existing literature by implementing a computationally efficient cyberbullying detection system using TFIDF feature extraction and supervised machine learning algorithms.

3.Proposed Methodology

The proposed methodology aims to develop an automated cyberbullying detection system using Machine Learning and Natural Language Processing techniques. The process begins with collecting a labeled dataset containing examples of both bullying and non-bullying text from publicly available sources such as social media datasets. The dataset serves as the foundation for supervised learning, where each text sample is tagged with its corresponding class label. Proper data selection ensures that the model learns meaningful patterns related to abusive and harmful language.

Once the dataset is collected, it undergoes a comprehensive preprocessing stage to remove noise and standardize the text. This step includes converting all text to lowercase, removing punctuation, URLs, special characters, and stop-words, and applying tokenization to break sentences into individual words. Lemmatization is performed to reduce words to their base form, thereby minimizing vocabulary size and improving model consistency. These preprocessing steps enhance the quality of input data and ensure better feature extraction.

After cleaning the text, TF-IDF (Term Frequency–Inverse Document Frequency) vectorization is applied to transform textual data into numerical feature vectors. TF-IDF assigns weights to words based on their importance within a document compared to the entire dataset. Words that appear frequently in a specific message but rarely across other documents receive higher weights, allowing the model to identify distinctive bullying-related terms. This transformation converts unstructured text into a structured mathematical representation suitable for machine learning algorithms.

Finally, the transformed dataset is split into training and testing sets to evaluate model performance. Supervised classification algorithms such as Support Vector Machine (SVM) and Multinomial Naive Bayes are trained on the feature matrix to learn patterns associated with cyberbullying. The trained model is then evaluated using performance metrics such as Accuracy, Precision, Recall, and F1-Score. This structured methodology ensures that the proposed system is accurate, scalable, and efficient for real-time cyberbullying detection.

3.1 Data Acquisition and Collection

The first step in the proposed system is the collection of a suitable and well-labelled dataset for cyberbullying detection. For this study, publicly available datasets from social media platforms such as Twitter and Kaggle repositories are considered. These datasets contain textual comments categorized into classes such as “bullying” and “non-bullying.” Proper labeling is essential because the system uses supervised machine learning algorithms that learn patterns based on predefined class labels.

The collected dataset typically consists of raw usergenerated text, which may include slang, abbreviations, emojis, URLs, and special characters. Since social media content is highly unstructured, the dataset must be examined carefully to ensure data quality and consistency. Any incomplete, duplicate, or irrelevant entries are removed to improve reliability and model performance.

In addition to data collection, class distribution analysis is performed to understand the balance between bullying and non-bullying samples. In most real-world datasets, neutral comments outnumber abusive ones, leading to class imbalance. Recognizing this imbalance at the initial stage helps in selecting appropriate evaluation metrics and improving model robustness. Proper data acquisition ensures that the foundation of the cyberbullying detection system is strong, diverse, and suitable for training accurate machine learning models.

3.2 Text Pre-processing and Data Cleaning

After data acquisition, the collected textual data undergoes a preprocessing stage to remove noise and standardize the content. Social media text is typically unstructured and contains irrelevant elements such as URLs, hashtags, mentions, emojis, punctuation marks, and special characters. These elements do not contribute meaningful information for classification and may negatively affect model performance. Therefore, they are removed using regular expressions and text-cleaning techniques.

The next step involves normalization of the text data. All characters are converted to lowercase to maintain uniformity and avoid duplication of words due to case differences. Stop-words such as “the,” “is,” “and,” and “are” are removed because they appear frequently but do not provide significant discriminatory information. Tokenization is then

performed to split sentences into individual words or tokens, enabling easier analysis and feature extraction.

Finally, lemmatization is applied to reduce words to their base or root form. For example, words like "running," "runs," and "ran" are converted to their base form "run." This reduces vocabulary size and ensures that similar words are treated as a single feature. Through these preprocessing steps, the raw textual data is transformed into a clean, consistent, and structured format suitable for TF-IDF feature extraction and machine learning classification.

3.3 Scope

After preprocessing the textual data, the next step is to convert the cleaned text into a numerical representation suitable for machine learning algorithms. Since machine learning models cannot directly process raw text, a feature extraction technique is required. In this project, Term Frequency–Inverse Document Frequency (TF-IDF) is used to transform textual data into weighted numerical feature vectors.

TF-IDF is a statistical technique that measures the importance of a word in a document relative to the entire dataset. The Term Frequency (TF) component calculates how frequently a word appears in a particular document, while the Inverse Document Frequency (IDF) component reduces the weight of commonly occurring words across all documents. The TF-IDF score is computed as:

$$TF\text{-}IDF = TF \times IDF$$

where

$$IDF = \log\left(\frac{N}{df}\right)$$

Here, N represents the total number of documents and df represents the number of documents containing the word. Words that appear frequently in a specific document but rarely in others receive higher weights, making them more significant for classification.

The TF-IDF vectorization process generates a feature matrix where each row represents a document and each column represents a unique word from the vocabulary. This matrix serves as the input for supervised machine learning algorithms such as Support Vector Machine (SVM) and Multinomial Naive Bayes. By emphasizing important bullyingrelated terms while minimizing common neutral words, TF-IDF enhances the model's ability to accurately distinguish between bullying and nonbullying content.

4. Objective

The primary objective of this study is to design and develop an automated cyberbullying detection system using Machine Learning techniques. The system aims to identify and classify harmful or abusive text content from social media platforms in an efficient and scalable manner. By automating the detection process, the project seeks to reduce reliance on manual moderation and improve online safety.

Another important objective is to collect and prepare a high-quality labeled dataset containing both bullying and non-bullying text samples. Proper data collection ensures that the system is trained on diverse examples of online communication. The dataset serves as the foundation for supervised learning and enables the model to understand patterns associated with harmful language.

The study also aims to implement effective Natural Language Processing (NLP) techniques for preprocessing and cleaning textual data. Since social media text often contains noise such as slang, emojis, special characters, and abbreviations, preprocessing steps like tokenization, stop-word removal, and lemmatization are necessary to improve data quality and model accuracy.

A key technical objective is to apply TF-IDF (Term Frequency–Inverse Document Frequency) feature extraction to convert textual content into numerical vectors. This method assigns importance to words based on their frequency and relevance, allowing the model to focus on significant bullying-related terms while reducing the impact of common words.

Another objective is to train and evaluate supervised machine learning algorithms such as Support Vector Machine (SVM) and Multinomial Naive Bayes for classification. The study aims to compare their performance and determine the most suitable algorithm for accurate and efficient cyberbullying detection.

Finally, the project aims to evaluate the overall system performance using metrics such as Accuracy, Precision, Recall, and F1-Score. By analyzing these metrics, the study ensures that the proposed system is reliable, balanced, and capable of detecting harmful content effectively in real-world scenarios.

5.Implementation

The implementation phase focuses on transforming the proposed methodology into a working cyberbullying detection system. The system is developed using Python because of its strong ecosystem for data science and machine learning applications. Libraries such as Pandas and NumPy are used for data manipulation and numerical computations, while Natural Language Toolkit (NLTK) is used for text preprocessing. Scikit-learn is utilized for TF-IDF vectorization, dataset splitting, and implementing classification algorithms.

The first step in implementation involves loading the collected dataset into the development environment and performing preprocessing operations. Regular expressions are applied to remove unwanted elements such as URLs, hashtags, user mentions, punctuation marks, and special symbols. The text is converted to lowercase to maintain uniformity, and stop-words are removed to eliminate frequently occurring but insignificant words. Tokenization and lemmatization are then applied to standardize word forms and reduce vocabulary size.

After preprocessing, the cleaned text is transformed into numerical format using the TfidfVectorizer from the Scikit-learn library. Important parameters such as maximum feature size and n-gram range are configured to optimize model performance. The vectorizer converts the textual data into a feature matrix where each document is represented as a weighted vector of word importance. This numerical representation serves as input for machine learning classifiers.

The dataset is then divided into training and testing sets using an 80:20 ratio. The training set is used to teach the model to recognize patterns associated with cyberbullying, while the testing set evaluates how well the model performs on unseen data. Supervised learning algorithms such as Support Vector Machine (SVM) and Multinomial Naive Bayes are trained using the training dataset.

Once the model is trained, predictions are generated for the testing data. The system's performance is evaluated using metrics such as Accuracy, Precision, Recall, and F1-Score. A confusion matrix is generated to analyze the number of correctly and incorrectly classified instances. These evaluation metrics help measure the reliability and effectiveness of the cyberbullying detection system.

Finally, the trained model and TF-IDF vectorizer are saved using job lib or pickle to enable real-time deployment without retraining. A simple user interface can be integrated to allow users to input text and receive instant classification results. The implementation ensures that the system is computationally efficient, scalable, and suitable for practical deployment in online communication platforms.

5.1 Environment Setup and Library Integration

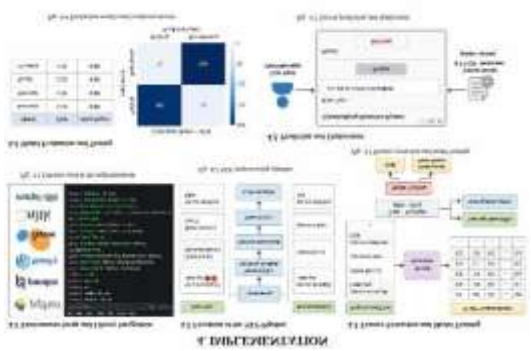
The implementation of the cyberbullying detection system is carried out using Python due to its extensive support for data analysis and machine learning libraries. The development environment includes tools such as Jupyter Notebook or an integrated development environment (IDE) like VS Code. Essential libraries such as Pandas and NumPy are used for data manipulation and numerical operations. NLTK (Natural Language Toolkit) is employed for text preprocessing tasks including tokenization and lemmatization. Scikit-learn serves as the primary library for implementing TF-IDF vectorization, dataset splitting, and classification algorithms. Matplotlib is used for visualizing performance metrics and confusion matrices.

5.2 Execution of the NLP Pipeline

The implementation begins with loading the dataset and executing preprocessing steps. Regular expressions are applied to remove URLs, mentions, special characters, and unnecessary symbols from the text. All text is converted to lowercase to maintain uniformity and avoid duplication of features due to case sensitivity. Stop-words are removed to eliminate commonly occurring words that do not contribute meaningful information to classification. Tokenization is performed to break text into individual words, followed by lemmatization to reduce words to their root form. This structured pipeline ensures clean and standardized input data.

5.3 Feature Extraction and Model Training

After preprocessing, the cleaned text is transformed into numerical form using the TfidfVectorizer from Scikit-learn. The vectorizer assigns weights to words based on their importance in the dataset. Parameters such as maximum features and n-gram range are tuned to enhance model performance. The dataset is then divided into training (80%) and testing (20%) sets. Supervised machine learning algorithms such as Support Vector Machine (SVM) and Multinomial Naive Bayes are trained on the training data to learn patterns associated with cyberbullying.



5.4 Model Evaluation and Testing

Once the model is trained, it is tested on unseen data to evaluate its performance. Predictions are generated for the testing dataset and compared with actual labels. Performance metrics such as Accuracy, Precision, Recall, and F1-Score are calculated to measure effectiveness. A confusion matrix is generated to analyze classification results, including true positives, true negatives, false positives, and false negatives. These metrics help assess the reliability and robustness of the system.

5.5 Prediction And Deployment

In the final stage, the trained model and TF-IDF vectorizer are saved using joblib or pickle for future use. This allows the system to make predictions without retraining the model each time. A simple user interface can be integrated where users input a text message, and the system instantly predicts whether it contains cyberbullying content. The implementation ensures that the system is computationally efficient and suitable for real-time deployment in social media environments.

6. Result And Discussions

The experimental results demonstrate that the proposed cyberbullying detection system performs effectively in classifying textual data into bullying and non-bullying categories. After applying TF-IDF vectorization and training supervised machine learning models, the system achieved high accuracy on the testing dataset. Among the evaluated classifiers, Support Vector Machine (SVM) showed better overall performance compared to Multinomial Naive Bayes in terms of precision and F1-score.

The confusion matrix analysis indicates that the majority of bullying instances were correctly identified as true positives, while most non-bullying texts were classified as true negatives. However, a small number of false positives and false negatives were observed. False positives occurred when neutral sentences contained strong words without harmful intent, while false negatives appeared in cases involving sarcasm or implicit abuse.

Performance metrics such as Accuracy, Precision, Recall, and F1-Score were used to evaluate the model comprehensively. High precision indicates that the system effectively minimizes incorrect bullying predictions, while good recall ensures that most harmful content is detected. The balanced F1-score confirms that the model maintains stability between precision and recall.

Overall, the results validate that TF-IDF combined with supervised machine learning algorithms provides a computationally efficient and reliable approach for cyberbullying detection. Although deep learning models may offer higher contextual understanding, the proposed approach achieves strong performance with lower computational complexity, making it suitable for real-time applications

7. Challenges And Limitations

One of the major challenges in cyberbullying detection is understanding context and sarcasm. TFIDF follows a bag-of-words approach and does not capture word order or deeper semantic relationships. As a result, sentences with hidden insults or sarcastic expressions may not be accurately classified.

Another limitation is dataset imbalance. In most realworld datasets, non-bullying comments significantly outnumber bullying comments. This imbalance can cause the model to become biased

toward the majority class, potentially reducing recall for bullying detection.

Language variation and evolving internet slang also pose challenges. Users often modify spellings or use coded language to bypass detection systems. Since TF-IDF relies on known vocabulary, new or unseen words may not be effectively recognized by the trained model.

Additionally, the system focuses only on text-based analysis and does not consider multimedia content such as images, videos, or audio. Cyberbullying can occur through memes or visual elements, which require advanced multimodal detection techniques beyond the scope of this study.

8.Conclusion

This research presents an automated cyberbullying detection system using TF-IDF feature extraction and supervised machine learning algorithms. The system effectively converts unstructured textual data into numerical form and applies classification techniques to identify harmful content.

Experimental evaluation shows that Support Vector Machine performs efficiently in detecting explicit abusive language, achieving high accuracy and balanced precision-recall values. The proposed model is computationally efficient and suitable for deployment in real-time social media monitoring systems.

Despite its limitations in handling sarcasm and contextual meaning, the study demonstrates that traditional machine learning approaches remain relevant for practical applications. The structured methodology ensures reproducibility and scalability.

In conclusion, the proposed framework contributes to improving online safety by providing a reliable and efficient cyberbullying detection mechanism. Future enhancements involving deep learning and contextual embeddings can further strengthen the system's detection capabilities.

9 Acknowledgement

We would like to express our sincere gratitude to our project guide, **Mr. Abdul Majeed** , Associate Professor, Department of Computer Science and Engineering (Data Science), Vidya Jyothi Institute of Technology, Hyderabad, for his valuable guidance, continuous support, and encouragement throughout the development of this project.

His insightful suggestions and motivation greatly contributed to the successful completion of this work. We would also like to thank the Head of the Department and faculty members of the CSE (Data Science)department for providing the necessary support and resources required for carrying out this project. We extend our sincere thanks to the Principal and management of Vidya Jyothi Institute of Technology for providing the infrastructure and academic environment that helped us complete this project successfully. Finally, we express our heartfelt gratitude to our parents, friends, and well-wishers for their constant encouragement and support during the course of this work.

9. Reference

[1] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," *Journal of Information Processing Systems*, vol. 14, no. 5, pp. 1026–1045, 2018.

[2] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying," in *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 2011, pp. 11–17.

[3] A. Rajaraman and J. D. Ullman, *Mining of Massive Datasets*. Cambridge University Press, 2011.

[4] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[5] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273– 297, 1995.

[6] T. Joachims, "Text categorization with Support Vector Machines: Learning with many relevant features," in *Proceedings of the European Conference on Machine Learning (ECML)*, 1998, pp. 137–142.

[7] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. O'Reilly Media, 2009.

[8] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.