IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 2, February 2026

DOI: 10.17148/IARJSET.2026.13225

Cyberbullying Detection on Social media Platforms using ML and NLP

Inbanathan S[1], Dr. P. Menaka[2]

Department of Information Technology, Dr. N.G.P. Arts and Science College, Coimbatore[1]

Associate Professor, Department of Information Technology, Dr. N.G.P. Arts and Science College, Coimbatore[2]

Abstract: The rapid growth of user-generated content on video-sharing platforms such as YouTube has significantly enhanced global communication while simultaneously increasing the prevalence of cyberbullying within comment sections. Offensive and harmful comments negatively impact users' psychological well-being and degrade online interactions. Manual moderation of such large-scale textual data is inefficient, necessitating automated intelligent detection systems. This study proposes a machine learning-based framework for multi-class classification of YouTube comments into bullying, non-bullying, and supportive categories. The system integrates Natural Language Processing (NLP) techniques including text cleaning, tokenization, stop-word removal, and lemmatization. Feature extraction is performed using TF-IDF vectorization with n-gram representations, along with sentiment-based features. To address class imbalance, SMOTE is applied during training. The classification framework employs XGBoost and Logistic Regression models, which are combined using a stacking ensemble approach to improve generalization and predictive performance. Experimental results demonstrate that the ensemble model effectively captures linguistic patterns in cyberbullying-related content and achieves reliable performance across accuracy, precision, recall, and F1-score. The proposed system provides a scalable and computationally efficient solution for automated cyberbullying detection on YouTube.

Keywords Cyberbullying Detection, YouTube Comment Analysis, Social Media Monitoring, Machine Learning, Natural Language Processing, Text Classification, Sentiment Analysis.

I. INTRODUCTION

The rapid expansion of social media platforms has transformed the way individuals communicate, share opinions, and engage with digital content. Among these platforms, YouTube stands out as one of the largest video-sharing services, hosting billions of videos and generating massive volumes of user interactions through comments. While this interactive ecosystem fosters creativity and global connectivity, it has also become a significant channel for cyberbullying, hate speech, harassment, and abusive behavior. Such harmful interactions can negatively impact users' mental health, self-esteem, and overall online experience.

Cyberbullying on YouTube commonly occurs in the comment sections of videos, where users can post offensive, threatening, or discriminatory remarks. The anonymity and accessibility of the platform often encourage toxic behavior, making manual moderation both time-consuming and insufficient. Traditional content moderation approaches rely heavily on manual review or keyword-based filtering, which are limited in detecting context-dependent, sarcastic, or subtly abusive language. As a result, harmful content may remain undetected, contributing to unsafe online environments. To address these challenges, automated cyberbullying detection systems based on Machine Learning (ML) and Natural Language Processing (NLP) techniques have emerged as effective solutions for large-scale comment analysis. By leveraging computational models, harmful content can be identified in real time with greater consistency, scalability, and efficiency compared to traditional moderation methods. Supervised learning algorithms enable the classification of textual data into predefined categories such as bullying, non-bullying, and supportive comments based on learned linguistic patterns and contextual information.

In this study, a robust machine learning framework is proposed that integrates advanced feature engineering and ensemble modeling techniques. Textual data is preprocessed using NLP methods including cleaning, tokenization, stop-word removal, and lemmatization. Feature extraction is performed using TF-IDF vectorization with n-gram representations, along with sentiment and linguistic features to enhance contextual understanding. To address class imbalance commonly observed in cyberbullying datasets, SMOTE is applied during training to improve minority class detection.

The classification framework employs XGBoost, and Logistic Regression as base classifiers, which are further combined using a stacking ensemble approach to enhance predictive performance and generalization capability. By leveraging the

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 2, February 2026

DOI: 10.17148/IARJSET.2026.13225

complementary strengths of multiple models, the proposed system effectively captures complex linguistic patterns in abusive language while maintaining computational efficiency. The performance of the proposed model is evaluated using standard multi-class classification metrics such as accuracy, precision, recall, and F1-score to ensure a comprehensive assessment of detection capability. By integrating optimized machine learning techniques with structured NLP processing, the system aims to provide an efficient, scalable, and reliable solution for automated cyberbullying detection on YouTube, thereby contributing to safer and more responsible online communication environments.

II. LITERATURE REVIEW

Several studies have explored automated cyberbullying detection using machine learning and natural language processing techniques across different social media platforms. M. Di Capua et al. [1] proposed an unsupervised approach combining textual and social features using Growing Hierarchical Self-Organizing Maps (GHSOM) and k-means clustering. Although promising results were achieved on the Formspring dataset, performance declined on YouTube and Twitter, highlighting challenges in cross-platform generalization.

J. Yadav et al. [2] implemented a transformer-based BERT model for cyberbullying detection, achieving high accuracy on Formspring and Wikipedia datasets. While effective in capturing contextual semantics, the approach required oversampling to handle data imbalance. Similarly, R. R. Dalvi et al. [3] focused on real-time Twitter data using TF-IDF with SVM and Naïve Bayes classifiers, where SVM achieved moderate accuracy, demonstrating the applicability of traditional machine learning models in real-time environments.

Trana R. E. et al. [4] analysed cyberbullying on YouTube using Naïve Bayes, SVM, and CNN models on a dataset of approximately 19,000 text instances. Their findings indicated varying performance across content categories, emphasizing the complexity of abusive language detection. N. Tsapatsoulis et al. [5] provided a comprehensive review of cyberbullying detection techniques, stressing the importance of integrating linguistic, behavioral, and contextual features for improved performance.

Furthermore, G. A. León-Paredes et al. [6] and P. K. Roy et al. [7] evaluated multiple machine learning and deep learning models, including Naïve Bayes, SVM, Random Forest, and Logistic Regression, reporting strong performance when appropriate NLP preprocessing techniques were applied. However, many existing studies focus primarily on Twitter datasets, with limited research specifically targeting YouTube comment analysis using ensemble-based approaches.

Based on the reviewed literature, there remains a need for a robust and scalable cyberbullying detection framework specifically tailored to YouTube comment analysis. The proposed study addresses this gap by leveraging advanced NLP preprocessing techniques and an ensemble-based machine learning approach to enhance detection accuracy, generalization capability, and computational efficiency.

Research gap: Most existing cyberbullying detection studies focus on Twitter datasets or computationally expensive deep learning models like BERT, with limited research specifically targeting YouTube comment analysis. Additionally, class imbalance and real-time scalability remain major challenges. Therefore, there is a need for a computationally efficient, scalable ensemble-based machine learning framework tailored for multi-class cyberbullying detection in YouTube comments.

III. PROBLEM STATEMENT

The rapid growth of user interactions on YouTube has led to an increasing presence of cyberbullying, abusive language, and harmful comments within video comment sections. Due to the massive volume of daily comments, manual moderation is inefficient and unable to ensure timely detection of inappropriate content. Existing keyword-based filtering methods often fail to capture context-dependent, sarcastic, or subtle forms of bullying, resulting in inaccurate classification and overlooked harmful behavior. Furthermore, class imbalance in real-world datasets makes accurate detection of bullying comments more challenging. Therefore, there is a need for an automated, scalable, and reliable machine learning-based system capable of accurately classifying YouTube comments into cyberbullying, non-bullying, and supportive categories to enhance online safety and improve the overall user experience.

IV. METHODOLOGY

The proposed system is designed to automatically classify YouTube comments into three categories: **Bullying, Non-Bullying, and Supportive Comments** using Machine Learning and Natural Language Processing (NLP) techniques.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 2, February 2026

DOI: 10.17148/IARJSET.2026.13225

The framework consists of five major stages: Data Collection, Text Preprocessing, Feature Engineering, Classification using Ensemble Learning, and Performance Evaluation.

1. Data Collection

YouTube comments are collected using the YouTube Data API or from an annotated dataset containing labeled comments. Each comment is manually categorized into one of the three classes: bullying, non-bullying, or supportive. Since cyberbullying datasets often suffer from class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) is applied during training to generate synthetic samples for the minority class and improve balanced learning. The dataset is divided into training and testing sets to evaluate model performance effectively.

2. Text Preprocessing

Raw YouTube comments typically contain noise such as emojis, URLs, special symbols, and inconsistent formatting. Therefore, a preprocessing stage is implemented to clean and standardize the data. All text is converted to lowercase to maintain uniformity. Punctuation, hyperlinks, and special characters are removed to eliminate irrelevant elements. The cleaned text is then tokenized into individual words. Stop words are removed to reduce unnecessary information, and lemmatization or stemming is applied to convert words into their base forms. These preprocessing techniques are fundamental steps in Natural Language Processing and are widely discussed in standard literature such as *Speech and Language Processing* by Jurafsky and Martin and *Natural Language Processing with Python* by Bird, Klein, and Loper. According to Manning and Schütze in *Foundations of Statistical Natural Language Processing*, normalization and token-level processing significantly improve feature consistency and reduce dimensionality in text classification tasks. By applying these established NLP preprocessing techniques, the proposed system enhances text consistency, reduces noise, and improves the overall quality of extracted features, ultimately leading to better classification performance.

3. Feature Extraction

After preprocessing, the textual data is transformed into structured numerical representations using **TF-IDF vectorization**, which assigns weights to words based on their importance within the dataset. To capture contextual information more effectively, **n-gram features** (unigrams and bigrams) are incorporated. Additionally, sentiment scores and linguistic features such as part-of-speech (POS) tags and rule-based indicators are included to enrich the feature space. This multi-dimensional feature engineering approach enables the model to better identify subtle linguistic patterns associated with bullying or supportive behavior.

4. Classification Using Random Forest

The classification framework employs XGBoost and Logistic Regression for cyberbullying detection. XGBoost effectively captures complex non-linear relationships in high-dimensional text data, while Logistic Regression provides stable linear decision boundaries and probabilistic outputs. To enhance predictive performance, a stacking ensemble approach is implemented, where predictions from the base models are combined using Logistic Regression as a meta-classifier to improve generalization and reduce bias. The theoretical foundation for ensemble learning and model combination is well established in *The Elements of Statistical Learning* by Hastie, Tibshirani, and Friedman, which highlights the effectiveness of boosting and stacking techniques in improving predictive accuracy.

5. Performance Evaluation

The proposed system is evaluated using multi-class classification metrics including accuracy, precision, recall, and F1-score. Accuracy measures overall classification correctness, while precision and recall evaluate the model's effectiveness in correctly identifying each class. The F1-score provides a balanced assessment, particularly important in imbalanced datasets. These evaluation metrics ensure a comprehensive analysis of the system's reliability and effectiveness in detecting cyberbullying and supportive behavior in YouTube comments.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 2, February 2026

DOI: 10.17148/IARJSET.2026.13225

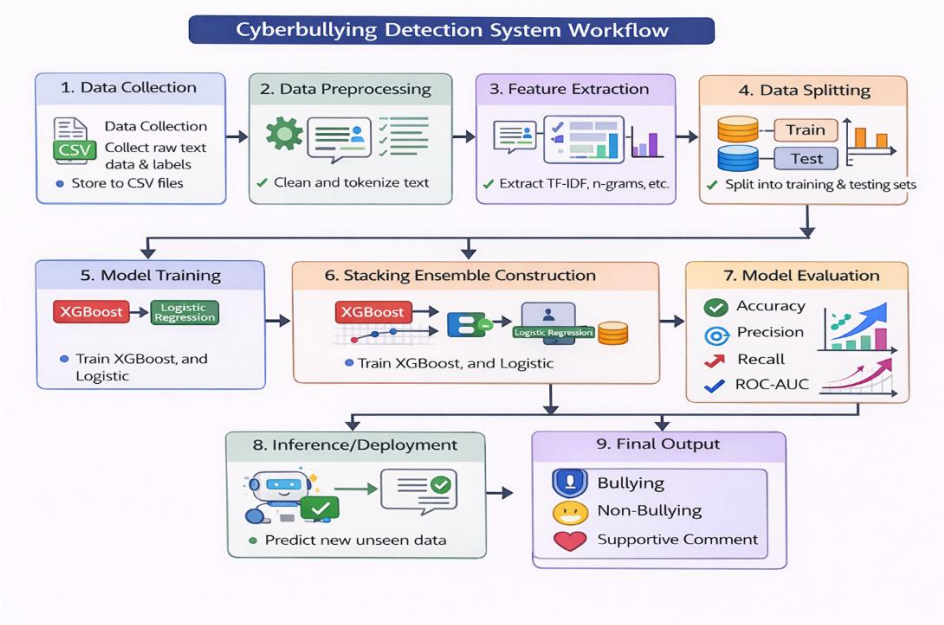

Fig:1.1: System Flow Diagram

V. PERFORMANCE AND EVALUATION

The performance of the proposed cyberbullying detection system is evaluated using standard multi-class classification metrics to ensure reliable and balanced assessment across all three categories: bullying, non-bullying, and supportive comments.

1. Evaluation Metrics

To measure the effectiveness of the Random Forest classifier, the following metrics are used:

Accuracy

Accuracy represents the overall correctness of the model in classifying YouTube comments.

$$\text{Accuracy} = \frac{\text{Total Correct Predictions}}{\text{Total Predictions}}$$

It provides a general measure of performance but may not fully reflect class imbalance.

Precision

Precision measures how many comments predicted as a specific class (e.g., bullying) are actually correct.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

High precision indicates fewer false positives.

Recall

Recall measures the model's ability to correctly identify all actual instances of a class.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

High recall ensures that most bullying comments are successfully detected.

F1-Score

The F1-score is the harmonic mean of precision and recall, providing a balanced evaluation.

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 2, February 2026

DOI: 10.17148/IARJSET.2026.13225

This metric is particularly important when dealing with multi-class classification problems.

Fig 1.2: Performance Evaluation

VI. RESULT AND DISCUSSION

The proposed cyberbullying detection system was evaluated using YouTube comments collected through the YouTube Data API, and the stacking ensemble model combining XGBoost and Logistic Regression achieved an overall accuracy of **77.58%**, demonstrating stable and reliable performance. The model produced balanced precision (0.7758), recall (0.7758), and F1-score (0.7758), indicating consistent classification across both Bullying and Supportive Comment categories. Class-wise results show strong detection capability for Bullying (precision: 0.7792, recall: 0.7684, F1-score: 0.7738) and Supportive Comments (precision: 0.7724, recall: 0.7831, F1-score: 0.7777), with the confusion matrix reflecting a well-distributed prediction pattern. The use of threshold tuning (0.65) further improved class balance and reduced bias, enhancing overall generalization. These results indicate that the proposed ensemble framework provides an effective and computationally efficient solution for automated cyberbullying detection on YouTube, while leaving scope for further refinement to handle subtle and context-dependent abusive expressions.

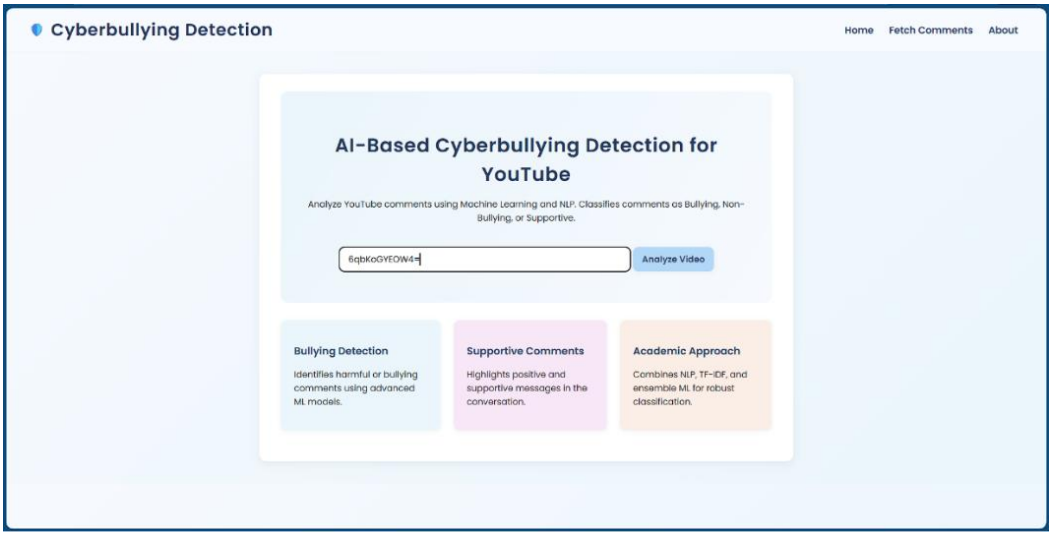

Fig 1.3: Home Page

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 2, February 2026

DOI: 10.17148/IARJSET.2026.13225

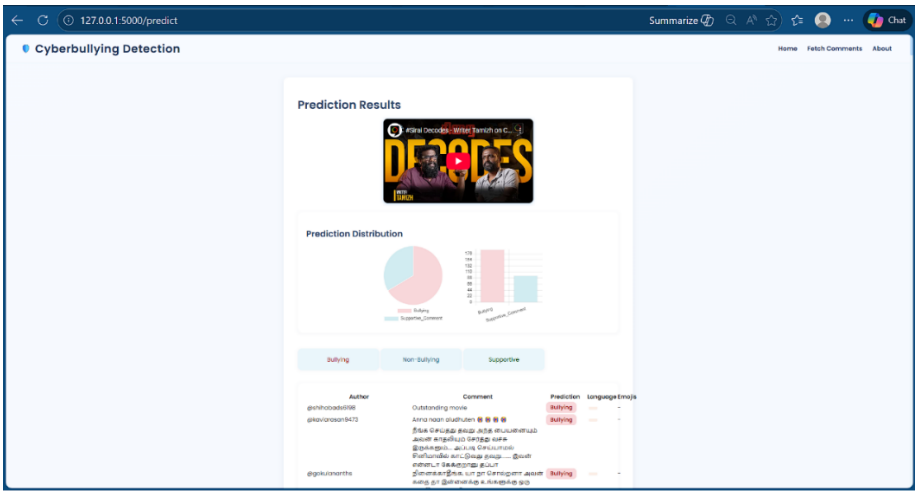

Fig 1.4: Prediction page

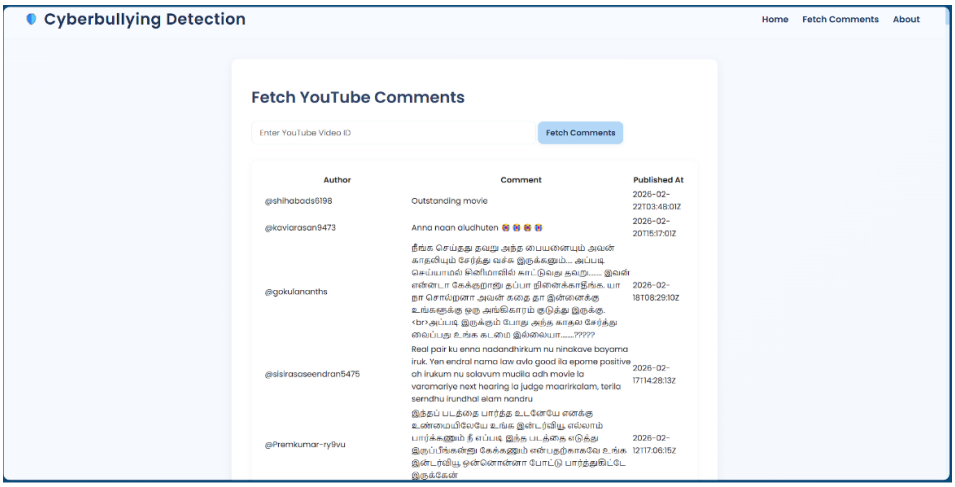

Fig 1.5: Fetching Page

VII. FUTURE WORK

Future work will focus on improving the recall of the bullying class by applying advanced class balancing techniques and cost-sensitive learning methods. Enhancing feature representation through semantic embeddings and incorporating slang, emoji, and sarcasm detection can further improve contextual understanding. Expanding the dataset to include multilingual comments and optimizing the system for real-time deployment are also important directions. These improvements can strengthen the robustness, scalability, and effectiveness of the proposed cyberbullying detection system.

VIII. CONCLUSION

This study presented a machine learning and NLP-based framework for detecting cyberbullying in YouTube comments using a stacking ensemble model. The results demonstrate that the proposed system provides a reliable and computationally efficient solution for automated moderation. Although challenges remain in detecting subtle abusive language, the system establishes a practical foundation for improving online safety on social media platforms. Future enhancements can further strengthen detection performance and real-time applicability.

REFERENCES

[1]. M. Di Capua, E. Petrosino, and A. Pescape, “Unsupervised cyberbullying detection in social media,” *IEEE Access*, vol. 8, pp. 204196–204209, 2020.

[2]. J. Yadav and D. K. Vishwakarma, “A BERT-based model for cyberbullying detection in social media,” *Procedia Computer Science*, vol. 167, pp. 1320–1329, 2020.

DOI: 10.17148/IARJSET.2026.13225

[3]. R. R. Dalvi, S. K. Patil, and M. A. Shaikh, "Real-time cyberbullying detection on Twitter using machine learning," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 4, pp. 389–395, 2020.

[4]. T. R. E. Trana et al., "Cyberbullying detection in YouTube comments using machine learning techniques," *Proceedings of the International Conference on Data Science and Applications*, 2019.

[5]. N. Tsapatsoulis, A. I. S. Panagiotopoulos, and G. P. Papadopoulos, "A review of cyberbullying detection methods in social media," *Information*, vol. 11, no. 9, pp. 1–20, 2020.

[6]. G. A. León-Paredes, M. L. Sánchez, and J. M. Paredes, "Cyberbullying detection in Spanish tweets using natural language processing techniques," *Applied Sciences*, vol. 11, no. 14, pp. 1–17, 2021.

[7]. P. K. Roy, J. S. Tripathy, T. K. Das, and X. Z. Gao, "A framework for hate speech detection using hybrid machine learning techniques," *Applied Sciences*, vol. 10, no. 22, pp. 1–22, 2020.

[8]. Jagtap, Y. D. Shelter Soul: Bridging Shelters and Adopters Through Technology. Shri Shivaji Vidya Prasarak Sanstha's Bapusaheb Shivajirao Deore College of Engineering, Dhule, India.

