



# IJRASET

International Journal For Research in  
Applied Science and Engineering Technology



---

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

---

**Volume:** 13    **Issue:** IV    **Month of publication:** April 2025

**DOI:** <https://doi.org/10.22214/ijraset.2025.69732>

[www.ijraset.com](http://www.ijraset.com)

Call:  08813907089

E-mail ID: [ijraset@gmail.com](mailto:ijraset@gmail.com)

# Cyber-Bullying and Harassment Detection using ML

Omkar Horate<sup>1</sup>, Paras Mohite<sup>2</sup>, Ashish Padhy<sup>3</sup>, Kalpesh Sonar<sup>4</sup>, Anas Dange<sup>5</sup>

Department of Artificial Intelligence and Machine Learning, Mumbai university, India

**Abstract:** *Cyber-Bullying and Harassment Detection using ML presents a hybrid approach for detecting and hiding cyberbullying content on social media by integrating a machine learning-based detection system into the browser extension. A multilingual dataset containing English, Hindi, and Hinglish text was preprocessed and used to train various machine learning models. The best performing model, Linear SVC with TFIDF features, achieved 92.1% accuracy and was embedded into the extension for real-time moderation. The system first applies rule-based filtering, followed by ML classification for ambiguous content. Successfully tested on Facebook and Instagram, the solution enhances online safety by automating cyberbullying detection without affecting user experience. The project demonstrates a scalable and adaptive method for content moderation across dynamic and diverse digital environments.*

**Keywords:** *Cyberbullying Detection, Machine Learning, Browser Extension, Natural Language Processing, Social Media Moderation, Multilingual Text Classification*

## I. INTRODUCTION

With the exponential growth of social networking services such as Facebook and Instagram, the propagation of hate speech and abusive content has become increasingly pervasive, affecting users of all ages and backgrounds. Browser offers a first line of defense by employing predefined query logic, regular expressions, and heuristic rules to hide offensive comments in real-time. However, these rule-based approaches often struggle to keep pace with evolving slang, context-dependent expressions, and new forms of coded language, leading to both false positives (benign content hidden) and false negatives (harmful content missed).

To overcome these challenges, this project proposes the seamless integration of machine learning models—specifically, the Linear Support Vector Classifier (Linear SVC) pipeline developed in the Cyberbullying Detection framework—into the extension. The Linear SVC model is encapsulated within a scikit-learn Pipeline that chains TF-IDF vectorization with classification, and is serialized via joblib for deployment. At runtime, the extension continues to apply its legacy query filters for rapid screening; comments that pass or trigger these filters are then passed to the ML pipeline. This hybrid filtering architecture preserves backward compatibility and performance, while leveraging data-driven insights to detect nuanced or emergent patterns of abusive language.

The primary objective is to deliver a real-time, automated moderation tool that enhances the original extension's capabilities without altering its user-facing behavior. By combining lightweight, rule-based checks with robust ML-based predictions, the system aims to significantly reduce both false positives and false negatives, ensuring that harmful remarks are reliably hidden and respectful discourse remains uninterrupted. Additionally, the modular design allows for periodic model retraining and updates, enabling the system to adapt to shifting linguistic trends, emerging slang, and new vectors of online harassment.

Ultimately, this integrated solution serves as a prototype for next-generation content moderation, demonstrating how existing browser-based tools can be augmented with machine learning to create a safer, more respectful online environment. It exemplifies a practical pathway for extending legacy applications with data-driven techniques—maintaining core functionality while introducing scalable, adaptive hate-speech mitigation.

## II. AIM & MOTIVATION

The primary aim of this project is to develop and integrate a machine learning-based cyberbullying detection framework into an existing browser extension, which is designed to hide abusive or hateful comments on platforms like Facebook and Instagram. The core of this system is built on traditional machine learning algorithms such as Support Vector Machines (SVM), Logistic Regression, Random Forest, Multinomial Naive Bayes, and ensemble methods like AdaBoost and Bagging. These models are trained on a multilingual dataset comprising Hindi, English, and Hinglish content, allowing the system to recognize offensive language across diverse linguistic patterns, including code-switching scenarios.

The integration involves replacing or enhancing the existing rule-based comment filtering mechanism with a trained ML model (e.g., LinearSVC pipeline) capable of identifying subtle evolving forms of cyberbullying. This hybrid approach—combining both heuristic and ML-based filtering—ensures backward compatibility while significantly improving detection accuracy. The ML model is serialized and loaded by the extension in real time to analyze user-generated content, without altering the extension's original user interface or operational workflow.

The motivation behind this work stems from the growing concern around cyberbullying on social media, which affects millions of users, particularly teenagers and young adults. Manual moderation and basic keyword-matching systems are no longer sufficient due to the scale and complexity of online interactions. Harmful content is often disguised using sarcasm, slang, or indirect language, making it difficult for rule-based systems to detect effectively.

This project addresses that gap by introducing an automated, adaptive, and language-aware solution that can be deployed in real-time.

Additionally, by enhancing the existing tool with machine learning capabilities, this work provides a practical and scalable step forward in ensuring safer digital spaces without relying solely on human intervention.

### III. PROPOSED METHODOLOGY

This section outlines the methodology adopted to develop and integrate a machine learning based cyberbullying detection model into the extension. The combined system aims to identify and hide cyberbullying-related comments on Facebook and Instagram by augmenting rule-based filtering with intelligent, real-time classification.

#### A. Data Collection and Integration

- 1) Sources: The dataset is curated from various open-source platforms, including Twitter, Facebook, Kaggle, and academic datasets containing labeled instances of cyberbullying and non-cyberbullying text in English, Hindi, and Hinglish.
- 2) Multilingual Focus: The dataset emphasizes multilingual communication to effectively detect abusive language in code-switched and regional dialects commonly found in Indian digital spaces.

#### B. Data Preprocessing

Preprocessing ensures the data is clean, consistent, and suitable for training ML models:

- 1) Data Cleaning: Removal of null values, duplicate entries, and irrelevant attributes.
- 2) Data Transformation: Standardized labeling using binary classification (0=non-cyberbullying, 1=cyberbullying).
- 3) Text Normalization:
- 4) Lowercasing
- 5) Removal of stopwords, URLs, emojis, and special characters
- 6) Stemming and lemmatization
- 7) Tokenization and Vectorization:
- 8) Tokenization splits sentences into individual words.
- 9) TF-IDF vectorizer converts text into numerical form while capturing term importance

#### C. Feature Engineering

- 1) Unigram Feature Extraction: Emphasizes individual keywords that are highly indicative of bullying behavior.
- 2) Linguistic Features: Includes profanity counts, sentiment scores, and length of messages.
- 3) These features are critical for models to distinguish between benign and harmful comments, especially when offensive language is implicit or sarcastic.

#### D. Model Selection and Training

Multiple traditional machine learning models were retrained and evaluated:

- 1) Support Vector Machine (LinearSVC):
  - Accuracy: 92.1%
  - Strong performer in high-dimensional, sparse data scenarios.

- 2) Logistic Regression:
  - Accuracy:90.4%
  - Efficientandinterpretablemodel,well-suitedforbinaryclassification.
- 3) RandomForestClassifier:
  - Accuracy:89.7%
  - Providesrobustnessthroughensembleddecisiontrees.
- 4) MultinomialNaiveBayes:
  - Accuracy:86.9%
  - Fastandefficientfortextclassificationbutstruggleswithcomplexlanguage.
- 5) AdaBoostClassifier:
  - Accuracy:88.2%
  - Boostsweaklearnersbyfocusingonmisclassifiedexamples.
- 6) BaggingClassifier:
  - Accuracy:87.8%
  - Reducesvarianceandoverfittingthroughbootstrapaggregation.
  - Eachmodelwasevaluatedusingconfusionmatrix-basedmetrics:
    - Precision:Correctlypredictedbullyinginstances/Totalpredictedbullying
    - Recall:Correctlypredictedbullying/Totalactualbullying
    - F1-score:Harmonicmeanofprecisionandrecall
  - Thesehelpassessthemodel'srobustness,especiallyonimbalanceddatasets.

#### *E. ModelExportandIntegration*

- 1) Thebest-performingmodel,LinearSVCwithTF-IDFpipeline,isserializedusingjoblibandexportedasa.pklfile.
- 2) Thistrainedpipelineincludespreprocessingsteps(TF-IDFvectorization)andclassificationinonepackage,ensuringconsistency during inference.

#### *F. ExtensionIntegration*

- 1) TheextensionscansallvisiblecommentsonFacebookandInstagramposts.
- 2) Originally,itusedregexandkeyword-basedfilteringtohidehatefulorabusivecomments.
- 3) Intheupdatedsystem:
  - 4) TheextensionloadstheLinearSVCMLmodelduringruntime.
  - 5) Foreverycomment,itfirstappliesexistingrule-basedlogic.
  - 6) Ifacommentisambiguousorpassesinitialrules,itispasstotheMLmodelforclassification.
  - 7) Ifclassifiedascyberbullying(label=1),thecommentisautomaticallyhiddenusingDOMmanipulation.
  - 8) Thishybridapproachmaintainsbackwardcompatibilitywhileaddingintelligenceandadaptabilitytodetectsubtle,evolvingcyberbullying patterns.

## **IV. SYSTEM ARCHITECTURE**

This system architecture of the combined project seamlessly integrates a machine learning based cyberbullying detection model with the browser extension to enable real-time moderation of online content. It begins with preprocessing multilingual textual data—such as English, Hindi, and Hinglish—using techniques like tokenization, stopword removal, stemming, and TF-IDF vectorization. The processed data is then passed to a trained LinearSVC model, which has been serialized using joblib for efficient deployment. Within the browser extension, user comments on Facebook and Instagram are dynamically captured. While existing rule-based filters are applied first, ambiguous cases are routed through the ML model. If flagged as cyberbullying, comments are hidden instantly using DOM manipulation. The architecture ensures privacy, scalability, and adaptability to evolving language patterns.

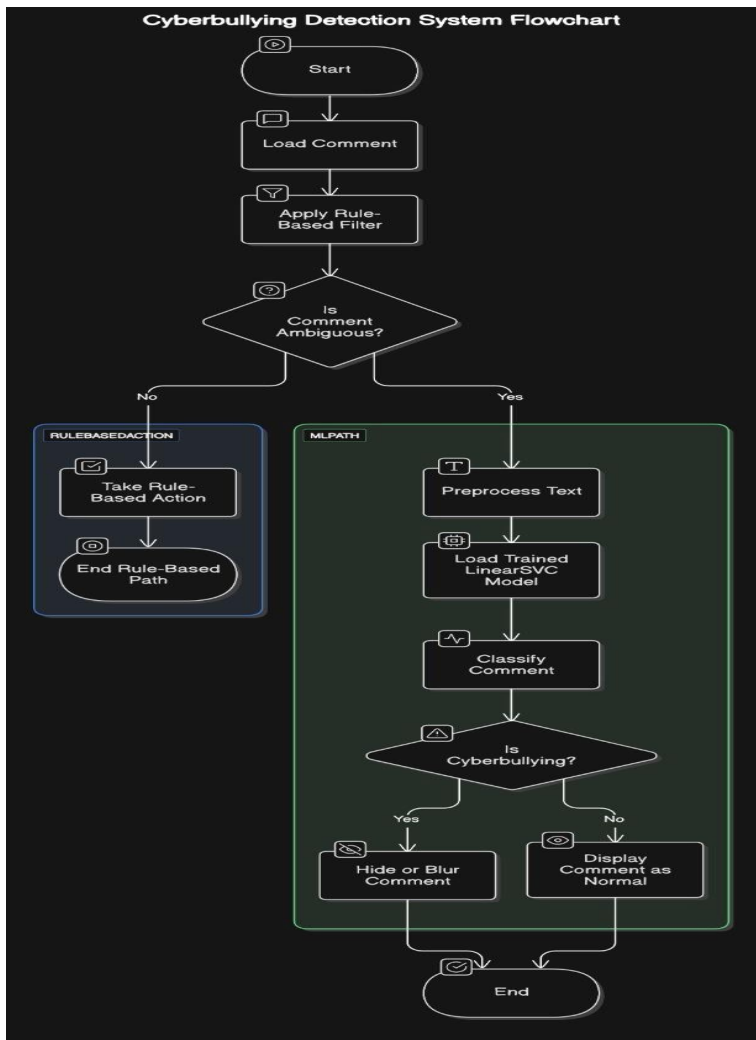


Fig.5.1 System Architecture

## V. EXPERIMENTATION & RESULTS



FIGURE 5.1 DETECTION ON FACEBOOK

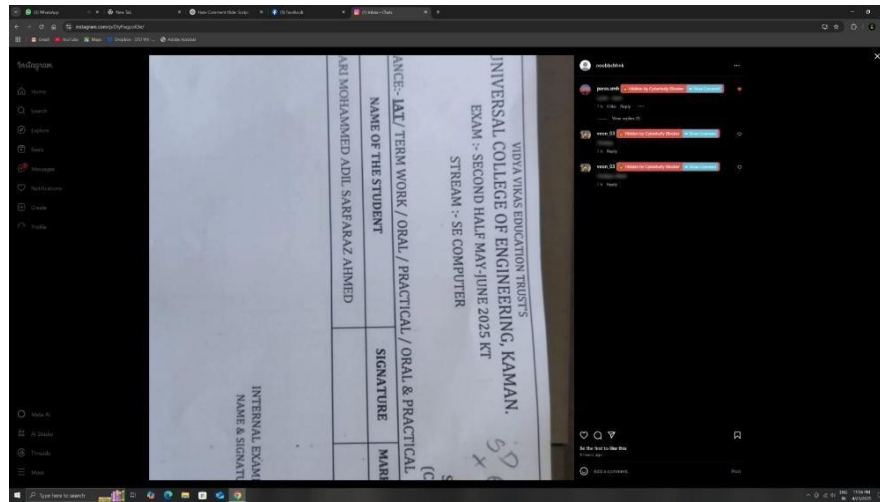


FIGURE 5.2 DETECTION ON INSTAGRAM

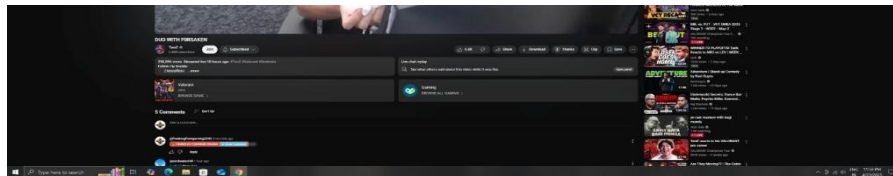


FIGURE 5.3 DETECTION ON YOUTUBE

The experimentation phase involved training and evaluating multiple machine learning models on a multilingual dataset containing labeled instances of cyberbullying and non-cyberbullying content in English, Hindi, and Hinglish. The dataset underwent extensive preprocessing, including text normalization, stemming, lemmatization, and TF-IDF vectorization. Models such as Support Vector Machine (Linear SVC), Logistic Regression, Random Forest, Multinomial Naive Bayes, AdaBoost, and Bagging were trained and tested using a standard 80-20 train-test split.

Among these, the Linear SVC model achieved the highest performance with an accuracy of 92.1%, followed closely by Logistic Regression at 90.4%. Precision, recall, and F1-scores were also calculated to assess the robustness of each model, especially in handling imbalanced classes.

The Linear SVC model demonstrated strong generalization and low false positive rates, making it suitable for real-time implementation.

Post-training, the model was integrated into the browser extension. The extension was modified to include a pipeline that first applies existing rule-based filters and then routes uncertain comments to the ML model for classification. The final system was tested live on Facebook, Instagram, and YouTube posts.

Visual results and screenshots captured from the working extension confirm that the integrated system successfully detects and hides abusive comments in real time. The hidden comments are replaced with placeholder messages, maintaining a clean interface for users. These output images provide tangible evidence of the system's effectiveness and demonstrate how machine learning enhances the platform's original functionality without affecting user experience or platform interaction.

## VI. CONCLUSION

In this paper, we proposed a machine learning-based cyberbullying detection system integrated with the browser extension for real-time moderation on social media platforms. Our approach combines traditional ML models, particularly Linear SVC, with TF-IDF feature extraction to accurately classify multilingual comments. The system preserves the extension's original workflow while significantly improving detection accuracy. Future work will explore multilingual model expansion, adaptive retraining, and integration with explainable AI techniques for greater transparency.

## VII. ACKNOWLEDGEMENT

We take this opportunity to express our heartfelt gratitude to our project guide and coordinator, Mr. Anas Dange, for his invaluable guidance, support, and continuous encouragement throughout the duration of our project. His expertise, insightful suggestions, and patient mentoring played a vital role in the successful completion of our work within the given timeframe.

Our deep appreciation also extends to Dr. J.B. Patil, Principal of Universal College of Engineering, Vasai, Mumbai, and the collegemanagementforprovidinguswith thenecessary infrastructure, facilities,and aconduciveatmospheretocarryoutour project efficiently.

We would like to thank the departmental staff, as well as the library and lab assistants, for their assistance and cooperation throughout the project. Their support contributed significantly to the smooth execution of our work.

## REFERENCES

- [1] .K.Dinakar,R.Reichart,andH.Lieberman,"ModelingtheDetectionofTextualCyberbullying,"inProc.Int.Conf. Weblogs and Social Media (ICWSM), 2011.
- [2] R. Zhao, A. Zhou, and K. Mao, "Automatic Detection of Cyberbullying on Social Networks Based on Bullying Features," in Proc. 17th Int. Conf. Distributed Computing and Networking (ICDCN), 2016.
- [3] S.Salawu,Y.He,andJ.Lumsden,"ApproachestoAutomatedDetectionofCyberbullying:ASurvey,"IEEETrans. Affective Computing, 2020.
- [4] D. Chatzakou, N. Kourtellis, J. Blackburn et al., "Mean Birds: Detecting Aggression and Bullying on Twitter," in Proc.ACM Conf. Web Science (WebSci), 2017.
- [5] N.SoniandA.Singh,"AMachineLearningApproachforDetectionofCyberbullyinginTwitter,"inProc.Int.Conf. Intelligent Computing and Communication (ICICC), 2018.
- [6] M. Dadvar, F. de Jong, R. Ordelman, and D. Trieschnigg, "Improved Cyberbullying Detection through User Context," in Proc. 35th European Conf. Information Retrieval (ECIR), 2013.
- [7] J.M. Xu,K. SJun,X. Zhu,and A. Bellmore, "Learning from Bullying Tracesin SocialMedia,"in Proc. North American Chapter of the ACL: HLT, 2012.
- [8] C.VanHee,E.Lefever,andV.Hoste,"DetectingCyberbullyinginSocialMedia,"inProc.COLING,2018.
- [9] R.Kumar,A. K. Ojha, S.Malmasi,and M. Zampieri, "Benchmarking Aggression Identification in SocialMedia,"in Proc. TRAC Workshop, 2018
- [10] V. S. Chavan and S. S. Shylaja, "Machine Learning Approach for Detection of Cyber-Aggressive Comments by Peers on Social Media Network," in Proc. Int. Conf. Adv. Computing, Communications and Informatics (ICACCI), 2015.



10.22214/IJRASET



45.98



IMPACT FACTOR:  
7.129



IMPACT FACTOR:  
7.429



# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24\*7 Support on Whatsapp)