International Journal of Advance and Applied Research
Peer Reviewed | International Open Access Journal
ISSN: 2347-7075 | Impact Factor – 8.26 | *Website:* https://ijaar.co.in/
Volume-14, Issue-1 | September - October 2026

Original Article

Cyberbullying Detection in Social Comments Using NLP

Krutika Sudhir Jadhav[1] & Chiranjiv Rakesh Churi[2]

[1]*M.Sc. Information Technology Part-I, SDSM College, University of Mumbai*

[2]*SDSM College, University of Mumbai*

Manuscript ID: IJAAR-140146

ISSN: 2347-7075
Impact Factor – 8.26

Volume – 14
Issue – 1
September- October 2026
Pp. 377 - 380

Submitted: 13 Oct 2026
Revised: 22 Oct 2026
Accepted: 25 Oct 2026
Published: 31 Oct 2026

Corresponding Author:
Krutika Sudhir Jadhav

Quick Response Code:

Website: https://ijaar.co.in/

DOI: 10.5281/zenodo.24366991

DOI Link: https://doi.org/10.5281/zenodo.24366991

Creative Commons

OPEN ACCESS

Abstract:

Nowadays internet and social media are an important part of our day-to-day life, as their use has grown rapidly in recent years. These applications provide the platform to the users where they can express their views, feelings and interact with each other. Cyberbullying refers to the usage of technology such as social media and messaging sites to hurt, humiliate or bully others. The Internet can sometimes be used to propagate abusive material and thus causing emotional and mental stress to individuals. Offensive and abusive use of language has become very prevalent on social media platforms and such activities cause serious psychological distress, depression and even suicide attempts among teenagers in particular. This project is focused on creating an automated tool for the prevention of cyberbullying on social networking sites through the use of Natural Language Processing and Machine Learning algorithms.

Creative Commons (CC BY-NC-SA 4.0)

This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0), *which permits others to remix, adapt, and build upon the work non-commercially, provided that appropriate credit is given and that any new creations are licensed under identical terms.*

How to cite this article:

Krutika Sudhir Jadhav & Chiranjiv Rakesh Churi (2026). Cyberbullying Detection in Social Comments Using NLP. International Journal of Advance and Applied Research, 14(1), 377 – 380. https://doi.org/10.5281/zenodo.24366991

Introduction:

In our day-to-day social media, it has become a very important means of communication. One of these activities is know as cyberbullying, which involves threatening or insulting other people through various online communications and platforms.

It has been found that 25% of the parents have been told by their children that they have become a victim of cyberbullying and

may receive psychological damages. The detection of cyber bullying models will entail the use of Natural Language Processing and Machine Learning technologies. Databases of various messages will be collected and various algorithms like logistic regression, Naive Bayes, SVM, and random forest will be used to detect bullying messages.

The term "cyberbullying" refers to the harassment, intimidation, or humiliation of individuals via the use of technology. In contrast with traditional forms of bullying, cyberbullying is persistent, pervasive, anonymous and can cause psychological damage to the victims. Victims can suffer from stress, anxiety, depression and even suicide.

It has been found that the current tradition of monitoring online content cannot cope with the amount of data generated every second. Hence, the requirement for intelligent automated systems that will help in detecting and preventing cyberbullying is essential. This is where Artificial Intelligence (AI) and Natural Language Processing (NLP) steps in.

Objectives:

Detection of cyberbullying in social comments using natural language processing tries to automatically identify and classify any negative content which is abusive or harmful to another person (insults, threats, harassment).

Key Objectives of NLP-Based Detection

1. Automated Real-Time Moderation: Scan vast amounts of user-generated comments instantly in order to lessen the burden of human moderators.
2. Preprocessing and Normalization of Text Data: Clean the unstructured text data of noise (URLs, special characters) and normalize slangs, abbreviations, emojis, sarcasm, and different languages (Hinglish).
3. Feature Extraction and Representation: extracting features in the form of vectors (numerical representation) from raw data by using one of the available
4. Methods: Bag of Words (CountVectorizer), TF-IDF, Word2Vec, FastText, or even contextual representations via transformers such as BERT.
5. Multi-Class Classification: classify the comments under a specific category (gender, religion, age, ethnicity or racial, or non-cyberbullying).
6. Investigation of Behavioral and Contextual Information: analyze the behavioral and contextual information of the users and their posts rather than blocking the keywords directly.

Literature Review:

"Cyberbullying Detection Using NLP and Machine Learning in Social Media Bi-Language".

Authors: Nikitha GS, Amritasri Shenoyy, K Chaturya, Latha JC, Janani Shree M. Author suggests the novel approach for detecting cyberbullying in English and Hindi languages by using NLP techniques and Machine Learning. The live tweets were collected by the searchers to perform this labeling study to analyze the relation of variables and cyber-aggression. Their model attempts to enhance the efficiency of cyber bullying detection by detecting abusive messages and classifying them according to the type of intimidation or threats. "Detection and classification of cyberbullying in social media by means of text mining" is the work dealing with detection and

classification of cyberbullying with the use of text mining. Authors: Not specified. Different machine learning classifiers are used for analyzing social media entries with an eye toward looking for patterns suggestive of bullying. Real-time analysis is underscored and emphasized as important along with requiring efficient algorithms against online harassment.

Research Methodology:

Data Collection:

- Sources: Extract data from social media sites such a Twitter,Face book, Instagram.
- Dataset: Utilize publicly available datasets (e.g., Kaggle's Cyber bullying dataset).
- Data Preprocessing: Text Cleaning Tokenization
- Feature Engineering: Bag of Words (BoW) TF-IDF (Term Frequency-Inverse Document Frequency).
- Model Selection: Machine Learning Models Deep Learning Models.
- Model Training & Validation: Accuracy, F1 Score, Recall, Precision

Results and Discussion:

Insight obtained from the sentiment analysis conducted to data collected from social media include : Sentiment analysis for social media posts about cyberbullying 70-80% of the undefined social media posts regarding cyberbullying have a negative statement, including fear, anger, and sadness 20-30% of neutral or empathic sentiment, indicating mixed option among the victims of cyberbullying, the detected emotions are distress and withdrawal, were as for bystander, the emotions of the outrange or indifference are detected Accuracy: Cyber detection Accuracy with Distil BERT 85-90% Accuracy Limitations: Decreased accuracy due to sarcasm detection. Differential accuracy is reported by machine learning models in the detection of cyber bullying: Traditional ML models (SVM, Naïve Bayes) → 65-75% accuracy Deep learning models (LSTM, CNN) → 75-85% accuracy BERT-based models (e.g., Ro BERT a Contextual meaning (e.g., "joking" vs. actual bullying) is difficult to classify. c. Perception Trends Among Users. User Group Perception & Behavior Victims Feel helpless, report anxiety and depression, but fail to report bullying because they are afraid of reprisal. Passersby 50% ignore, 30% assist, 20% engage passively (likes, shares). Offenders Some excuse bullying as a prank, others have no idea about its affected. Platform-Specific Insights Twitter/X → Most hate speech is caused by harassment and trolling. Instagram → Indirect bullying is caused by picture taunts and ostracism. Facebook → Group bullying and disinformation are responsible for cases of cyber bullying. Effectiveness of NLP-Based Interventions Automated Moderation: 60- 70% of the content of cyber bullying is detected by AI filters, but false positives remain a problem. Sentimentaware chat bots: Support victims to reach mental health adoption is low. NLP models for early warnings: They classify cyberbullying 80%.

Discussion:

a) NLP models in Cyberbullying Perceptions Perception of NLP models is used to classify cyberbullying cases and public opinion. They do not do well with sarcasm, coded messages, and evolving slang. Improved context-aware models (e.g., Chat GPT, GPT-4, Ro BERT a are enhancing detection accuracy.

b) Ethical Issues and Bias in AI Assisted Identification There may be cases where some non-threatening statements are flagged as being bullying and blocked.
c) Future Improvements: Rule-Based Approach and Deep Learning Approach: Both of these approaches will increase detection accuracy.
d) Real-Time Monitoring: AI based real-time cyber bullying detection would allow intervention before it escalates.
e) Personalized Support Systems: Individualized AI-powered mental health facilities especially for the victims can work better during intervention.

Conclusion:

Cyberbullying is one of the significant problems in the current digital world. This paper proposes a method to detect it using natural language processing approaches. Various machine learning models have been taken into consideration in order to analyze which among them demonstrates better performance for the task of text classification. The whole experiment was carried out on a global Twitter dataset. LR possesses the maximum value of accuracy among all models which clearly indicates its better performance in comparison with others. It has been also observed that the accuracy and speed of prediction by Logistic Regression increases with the increase in the data size.

References:

1. Aditya Desai, Shashank Kalaskar, Omkar Kumbhar, Rashmi Dhumal. "Cyberbullying Detection on social media using Machine Learning", ITM Web of Conferences 40, 03038, 2021.
2. Fawzya Ramadan Sayed, Eman Hassan Elnashar, Fatma A Omara. "Cyberbullying Detection in social media Using Natural Language Processing", Scientific African, e02713, 2025.
3. Stephen Afrifa, Vijayakumar Varadarajan. "Cyberbullying Detection on Twitter using Natural Language Processing and Machine Learning Techniques", International Jornal of Innovative Technology and Interdisciplinary Sciences 5 (4), 1069-1080, 2022.
4. Andrea Perera, Pumudu Fernando. Accurate "Cyberbullying Detection and Prevention on Social Media", Procedia Computer Science 181, 605-611, 2021.
5. Diva, D., Aherkar, S., Choudhary, M., Nair, S., & Patil, C. (2026). Cyberbullying Detection in social media using NLP. International Journal of Creative Research Thoughts, 14(4), j555-j561. https://ijcrt.org/viewpaperforall.php?paper=IJCRT26A4111