# Detection of Cyberbullying On Social Media Using Machine Learning

[1]Prof.Vishal Shinde, [2]Ms. Renuka Vikas Hase , [3]Ms. Priti Pandharinath Chaudhari , [4]Ms. Nayna Dattaram Vashiwale

[1]Asst.Professor,[2,3,4]UG Student,[1,2,3,4]Computer Engg. Dept. Shivajirao S. Jondhle College of Engineering & Technology, Asangaon, Maharashtra, India.

[1]mailme.vishalshinde@gmail.com, [2]renuhase9102@gmail.com, [3]pritic757@gmail.com, [4]naynavashiwale@gmail.com

**Abstract - Cyberbullying is one of the biggest problems on the internet, affecting adults and teens alike. It has led to suicide and sadness. The need for content regulation on social media platforms is growing. research that follows makes use of natural language processing (NLP) and machine learning (ML) to create a model to detect cyberbullying within text data. The data comes from two different categories, hate speech tweets on Twitter and comments on personal assaults on Wikipedia forums. The study looks at three different feature extraction techniques, as well as four classifiers, to determine the best strategy. The model achieves accuracy levels of more than 90% on Twitter data and more than 80% on Wikipedia data.[1]**

**Keywords: Cyberbullying, natural language processing, Machine Learning (ML), Classifiers, Social *Media.***

## I. INTRODUCTION

More than ever, technology has ingrained itself into our daily lives. as the internet has developed. These days, social media is in style. However, just like with everything else, there will undoubtedly be misusers, they may appear early or late. These days, cyberbullying is widespread. Social networking sites are great resources for internal communication. Social networking use has grown throughout time, but generally speaking, people discover unethical and immoral ways to do bad things. This is observed occasionally occurring amoungst young adults or teenagers. One of their harmful behaviour is cyberbullying one another. Cyberbullying is when someone is harassed, threatened, embarrassed, or targeted for amusement or even with deliberate intent through well-thought-out methods. [1]Sometimes users post harmful or suicide-related tweets on their account on twitter user's text messages to their friend about making suicide so its main work to detect and identify that post of users. Admin keeps watch on users posts so that users cannot do anything wrong with their life.[7]

## II.AIMS & OBJECTIVE

### a)Aim

The primary goal of this research is to employ machine learning to detect language patterns used by threats and to create standards for identifying onlineharassment. English has been the languageused in the majority of studies on machine learning-based cyberbullying detection.

### b) Objective

The aim of this document is to categorizetextual information as either hate speech ornon-hate speech. If the method is found to be hate speech, it may also identify the hatespeech's targeting characteristics, such as hate speech related to race.

## III.LITERATURE SURVEY

**Paper 1: Automatic Detection of Cyberbullying on Social Networks basedon Bullying Features.**

Cyberbullying behaviour has drawn moreand more attention as social media use rises. Cyberbullying may result in teen suicide among other grave and detrimental effects on a person's life. One practical way to lessen and end cyberbullying is to use proper machine learning and natural language processing algorithms to automatically identify bullying content. Buta lot of the literature's current methods are essentially standard text classification models that don't take bullying traits into account.introduce a representation learning system designed specifically for the detection of cyberbullying in this study.[4]

**Paper 2: Collaborative of cyberbullying behaviour in twitter data.**

Unwanted user behaviour on Twitter are growing along with the platform's data size. Cyberbullying is one of these unwanted behaviours that might have disastrous results. Therefore, it is imperative to effectively recognize occurrences of cyberbullying by real-time tweet analysis. Common methods for identifying cyberbullying are primarily isolated, which makes them time-consuming.

DOI : 10.35291/2454-9150.2024.0264

Through the application of collaborative computing techniques, this research enhances thedetection task. In this study, many collaboration paradigms are proposed and discussed. According to preliminary findings, the detection process is faster andmore accurate than it was with the stand-alone paradigm.

The tweets in a distributed and cooperative manner in order to locate cyberbullying incidents in textual data from Twitter. The tweets are categorized into cyberbullying various machine learning algorithms by this collaborative paradigm.[5]

**Paper 3: Supervised machine learning for the detection of troll profiles in twitter social network: application to a genuine cyberbullying incident.**

Users now have the capacity to maintain anonymity thanks to the adoption of new technologies and the success of social networks. No one can verify whether a profile and a real person match because it is possible to construct an alter ego that has no connection to the original user. This issue leads to daily scenarios when individuals with fictitious profiles, or at least ones unrelated to their true identities, post news, reviews, or multimedia content in an effort to malign or attack others who could or might not be aware of the attack. These behaviours may have a major effect on the surroundings of the victims they affect, creating scenarios where virtual attacks turn into deadly real-world event.[6]

**Paper 3: Detection of Suicide Related Posts in Twitter Data Stream.**

Suicidal ideation detection in online social networks is an emerging research area with major challenges. Recent research has shown that the publicly available information, spread across social media platforms, holds valuable indicators for effectively detecting individuals with suicidal intentions. The key challenge of suicide prevention is understanding and detecting the complex risk factors and warning signs that may precipitate the event. In this paper, present a new approach that uses the social media platform Twitter to quantify suicide warning signs for individuals and to detect posts containing suicide-related content. The main originality of this approach is the automatic identification of sudden changes in a user's online behaviour. To detect such changes, its combine natural language processing techniques to aggregate behavioural and textual features and pass these features through a martingale framework, which is widely used for change detection in data streams.[7]

## IV. EXISTING SYSTEM

I-Hsien Ting was awarded a A approach that included opinion mining, social network analysis, and keyword matching generated recall of 0.71 and precision of 0.79 using datasets from four websites. [2]

E. Wulczyn, N. Thain, and L. Dixon, has approached Platforms combat this with policies concerning such behaviourWikipedia has a policy of "Do not make personal attacks anywhere in Wikipedia" and notes that attacks may be removed and the users who wrote them blocked.[4]

In order to create results, data and results from many detection nodes each of which utilizes a different or same algorithm are integrated in a collaborative detection approach that was proposed by Mangaonkar et al.[5]

Using the Long Short-Term Memory (BLSTM) model that is bidirectional, Kai Eckert trained deep neural networks on Twitter, Wikipedia, and Form spring datasets. He then applied the model to the YouTube dataset and obtained an F1 score of 0.97.[6]

## V. COMPARATIVE STUDY

*Table.1: Comparative Analysis*

| Sr.no | Author | Paper Title | Purpose | Technology |
|---|---|---|---|---|
| **1.** | Varun jain, Vishant Kumar And Vivek Pal | Detection of Cyberbullying on Social Media UsingMachine learning | Natural Language Processing & Machine Learning | The study that follows makes use of data from two distinct forms of cyberbullying: hate speech-related Twitter tweets and personal attack-based Wikipedia forum comments. |
| **2.** | Rui Zhao, Anna Zhou & Kezhi Mao | Automatic Detectionof Cyberbullying on Social Networks based on Bullying Features. | Natural Language Processing & Machine Learning | Furthermore, it also provide an effective real-world application example where this methodology was used to identify and put an end to cyberbullying. |
| **3.** | Amrita Mangaonkar, Allenoush Hayrapetian, RajeevRaje | Collaborative of cyberbullyingbehaviour in twitter data. | Machine Learning | If at all feasible, it is imperative to detect incidents of cyberbullying immediately through the analysis of tweets. |
| **4.** | Patxi galan-garica, carlos laorden, jose Gaviria de la Puerta, lgos santos | Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying. | Machine Learning | It is impossible to verify whether a profile accurately represents the genuine user because it is easy to construct an alter ego that is unrelated to the original person. |

## VI. PROBLEM STATEMENT

The increased use of social networking and the freedom of speech have provided people ofall demographics with the best environmentfor cyberbullying and cyber aggression. This has substantial and observable effects on the victim's conduct, ranging from emotional distress and social isolation to more serious and fatal outcomes. The goal of automatically detecting cyberbullying has proven to be extremely difficult becausesocial media content is typically provided in an unstructured free-text format, devoid of linguistic conventions, regulations, and standards. As was previously said in the literature review section, it appears that a sizable number of research studies are primarily concerned with identifying textual patterns associated with cyberbullying across various social media platforms. Nevertheless, the majority of automated methods and detection strategiesdeveloped are for resource-rich

## VII. PROPOSED SYSTEM

The ensuing work uses machine learning and natural language processing to develop a model based on the identification of cyberbullying in text data. The study makes use of data from two distinct types of cyberbullying: hate speech-related Twitter tweets and personal attack-based Wikipedia forum comments.

**MODULES:**

- User
- Admin
- Data Preprocessing
- Machine Learning

## VIII. ALGORITHM

**Operational procedures inside the machine learning architecture :-**

Step 1: Bring in the necessary Python libraries

Step 2: Examine the data file.

Step 3: Clean up the data.

Step 4: Use Count Vectorizer to transform text into vector form.

Step 5: Divide the test and train samples from the dataset.

Step 6: Extract features using tf-idf.

Step 7. Use the collected features to train ML classifiers (SVM, LR, RF, NLP classifier).

Step 8: Use test samples to evaluate the learned models.

**Logistic Regression**

Set up an object for a logistic regression model: Logistic Regression (lgr)

Utilizing the oversampled data, fit the model:X_over: Features in input data that are oversampledy_over: Target labels with data that has been oversampled: lgr.fit(X_over, y_over)

Using the test data as well as the proficient model, forecast:
X_test: Test information characteristics
lgr.predict(X_test) ← y_pred

Produce a report on classification:

True labels are obtained based on the test results using the following formula: lg_cr←classification_report(y_test, y_pred, output_dict←True)

## XI. MATHEMATICAL MODEL

**1. Feature extraction:** Using methods like word embeddings and TF-IDF, the text input is transformed into the feature matrix

\(X\). A social media post is indicated by each row of \(X\), and a word or phrasefrom the post is represented by each column's feature.

**The Logistic Regression Model:** This model is employed in binary classification.The following is the model's prediction forthe likelihood that a specific societal media

2. constitutes cyberbullying (\(Y = 1\)):

= \frac{1}{1 + e^{-(X\theta)}}\[P(Y=1|X,

\theta)]

where the natural logarithm's base isrepresented by \(e\).

**3. Training the Model:** In this step, the logistic loss (cross-entropy) is minimizedin relation to the model parameters

\(\theta\):

\[J(\theta) = -\frac{1}{m}

\sum_{i=1}^{m}\left[y^{(i)}\log(h_{\theta}(x^{(i)})

where \(h_{\theta}(x^{(i)})\) is the expected probability for the \(i\)-th trainingexample and \(m\) is the number of training examples.

## XII. SYSTEM ARCHITECTURE

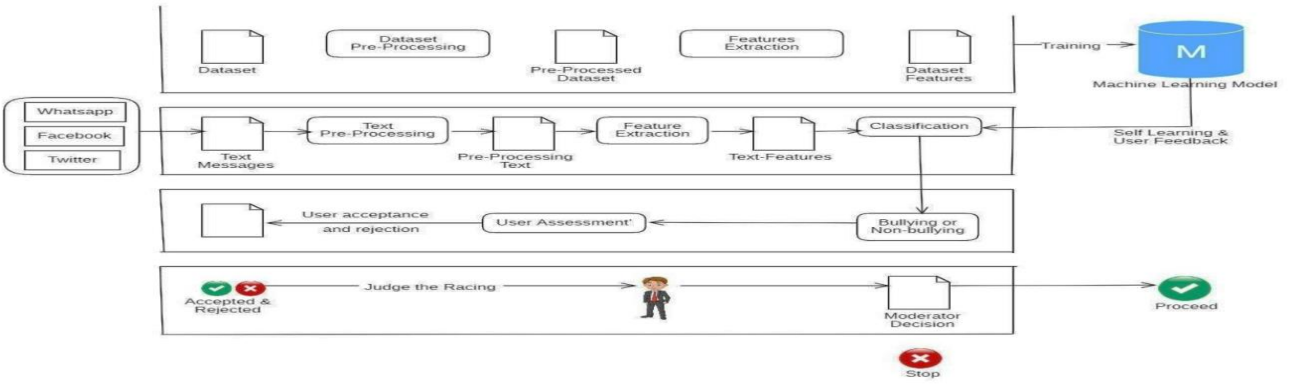


*Fig.1 System Architecture*

DOI : 10.35291/2454-9150.2024.0264

Data is collected from a range of social media sites, including WhatsApp, Facebook, Twitter, and text messages. Dataset Pre-Processing Raw data is cleaned, organized, and transformed into a suitable format for analysis. This may involve tasks such as eliminating copies, managing absent values, and standardizing the data Text Pre-Processing Textual data undergoes further processing to get it ready for examination. This includes tokenization, lowercasing, punctuation removal, and possibly stemming or lemmatization.

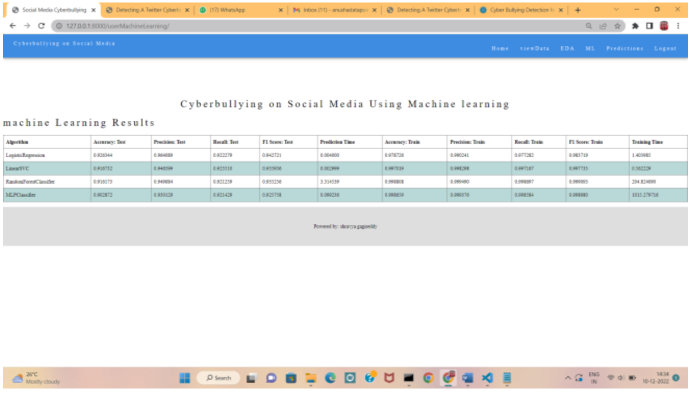
Feature Extraction Relevant features are extracted from the pre-processed data. These features could include linguistic patterns, sentiment analysis, word frequencies, and other indicators of cyberbullying behaviour. User Assessment and Feedback, User feedback may be incorporated into the system to improve its performance. This could involve manual assessment of labeled data or feedback mechanisms for users to report instances of cyberbullying. Machine Learning model is trained on the extracted features to classify instances of cyberbullying. This could involve various algorithms such as Support Vector Machines (SVM), Random Forests, or NLP ClassifiersClassification The trained model is used to classify new instances of text data as either cyberbullying or non-cyberbullying.

Self-Learning where it continuously improves its performance over time based on new data and feedback.Moderator Decision In some cases, a human moderator may be involved in the decision-making process, particularly for ambiguous cases or to handle false positives/negatives.

## XI. ADVANTAGES

- Early Intervention: By recognizing cyberbullying, victims can receive prompt assistance to stop further harm from occurring.
- Safety: Cyberbullying helps to promote a safer online environmentby deterring potential bullies andpromoting a more positive online community.
- Education: Parents and schools can utilize these detections as opportunities to teach kids about appropriate behaviour on the internet.
- Privacy Protection: Guards users' data and privacy againstexploitation or harassment.

## XI. SCREENSHOTS


Social Media Cyberbullying
Detecting A Twitter Cyberb
(17) WhatsApp
Inbox (11) - anushadadapo
Detecting A Twitter Cyberb
Cyber Bullying Detection
127.0.0.1:8000/userMachineLearning/
Cyberbullying on Social Media
Home
viewData
EDA
ML
Predictions
Logout

Cyberbullying on Social Media Using Machine learning

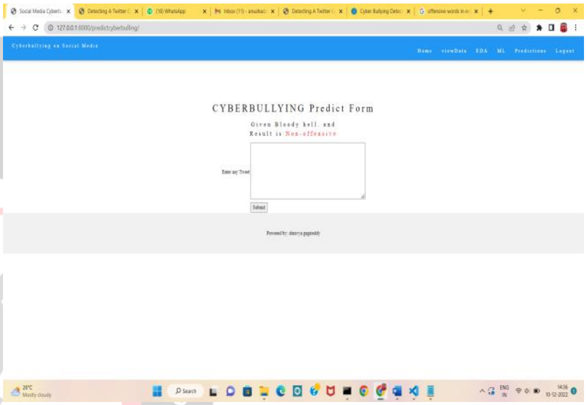
machine Learning Results

| Algorithm | Accuracy: Test | Precision: Test | Recall: Test | F1 Score: Test | Prediction Time | Accuracy: Train | Precision: Train | Recall: Train | F1 Score: Train | Training Time |
|---|---|---|---|---|---|---|---|---|---|---|
| LogisticRegression | 0.936344 | 0.964089 | 0.922279 | 0.942721 | 0.004800 | 0.978736 | 0.980241 | 0.977282 | 0.983719 | 1.405985 |
| LinearSVC | 0.916732 | 0.946599 | 0.925510 | 0.935936 | 0.002999 | 0.997019 | 0.998298 | 0.997167 | 0.997735 | 0.562229 |
| RandomForestClassifier | 0.916173 | 0.949884 | 0.921259 | 0.935256 | 3.314539 | 0.998808 | 0.999480 | 0.998687 | 0.999085 | 204.824890 |
| MLPClassifier | 0.902872 | 0.930129 | 0.921429 | 0.925758 | 0.089236 | 0.998859 | 0.989376 | 0.998584 | 0.998880 | 1015.278716 |

Powered by: sharvya gaganeddy

*Fig 2. Machine Learning Results*

In this fig we can see the machine learning result in which having the 4 algorithm LR, LinearSVC, Random forest classifier and MLP classifier, but In which Logistic Regression giving the most correct accuracy in results.



CYBERBULLYING Predict Form

Given Bloody hell. and
Result is Non-offensive

Enter any Tweet

Submit

Powered by: sharvya gaganeddy

*Fig 3. Prediction Form*

In this prediction form we can see the tweet or any message is offensive or non-offensive according our filled information in the form.

## XIII.CONCLUSION

Thus, we have tried to implement the paper, "Detection of Cyberbullying on Social Media UsingMachine learning" by "Varun jain, Vishant Kumar And Vivek Pal V, 2021 and the conclusion is as follows : we have successfully executed the machine learned classifier algorithms to test web information contents. In this paper we successfully presented an architecture for cyberbullying detection in this study. We spoke about the architecture for two different kinds of data: personal attacks on Wikipedia and hate speech data on Twitter. Because hate speech is often associated with profanity, natural language processing approaches have shown promising results in detecting hate speech with accuracy rates of over 90% when utilizing basic machine learning algorithms.

## REFERENCES

[1] Varun Jain , Vishant Kumar & Vivek Pal,"Detection of

DOI : 10.35291/2454-9150.2024.0264

Cyberbullying on Social Media Using Machine learning", 2021. https://doi.org/10.1109/ICCMC51019.2021.9418254

[2] P. I. H. Ting, W. S. Liou, D. Liberona,S. L. Wang, and G. M. T. Bermudez, "Towards the detection of cyberbullying based on social network mining techniques," in Proceedings of 4th International Conference on behavioural, Economic, and SocioCultural Computing, BESC 2017, 2017, vol. 2018-January, doi: 10.1109/BESC.2017.8256403.

[3]M. Dadvar and K. Eckert, "Cyberbullying Detection in Social Networks Using Deep Learning Based Models; A Reproducibility Study," arXiv. 2018.

[4] E. Wulczyn, N. Thain, and L. Dixon, "Ex machina: Personal attacks seen at scale,"2017,doi:10.1145/3038912.305259.

[5] Rui Zhao, Anna Zhou & Kezhi Mao," Automatic Detection of Cyberbullying on Social Networks based on Bullying Features",2016.https://doi.org/10.1145/2833312.28495671

[6] A. Mangaonkar, A. Hayrapetian, and R. Raje, "Collaborative detection of cyberbullying behaviour in Twitter data,"2015, doi: 10.1109/EIT.2015.7293405.

[7] Patxi galan-garica, carlos laorden, jose Gaviria de la Puerta, lgos santos "Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying.",2014. DOI: 10.1007/978-3-319-01854-6_43

[8] Mr. Vishal Shinde, Mr. Vaibhav Khare, Mr. Atul Patil, Mr.Pravin Shinde, "Detection of Suicide Related Posts in Twitter Data Stream",2020.DOI:10.35291/24549150.2020.0219.