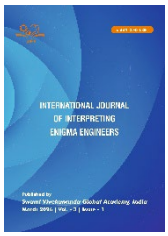
International Journal of Interpreting Enigma Engineers (IJIEE)
https://ejournal.svgacademy.org/index.php/ijiee/
https://doi.org/10.62674/ijiee.2025.v3i01.001
eISSN: 3048-6300
OPEN ACCESS

AN ENHANCED MACHINE LEARNING FRAMEWORK FOR CYBERBULLYING DETECTION IN SOCIAL MEDIA TEXT MESSAGES



Nithya Lakshmi N*, K Akhil, R Mahesh, L Sai Praneeth

Original Article

MVSR Engineering College, Hyderabad-501510, India

*Corresponding Author's Email: nithya_it@mvsrec.edu.in, 245120737135@mvsrec.edu.in, 245120737133@mvsrec.edu.in, 245120737138@mvsrec.edu.in

Abstract

Abstract: The rapid usage of internet makes it easy for the people to communicate across the globe and use the social media. Cyberbullying is considered to be a form of online harassment that creates harsh consequences like mental health issues, social isolation and even suicide. Nowadays Cyber bullying on social media has become a widespread issue in digital age that causes harmful and negative impacts on people. This paper mainly focuses on finding those cyberbullying messages employing language processing methods thereby processing textual data and few machine learning algorithms to characterize them and detect correlating them with the preprocessed data. From the observations, it is noted that when compared to all the other machine learning used the gradient descent method seems to perform better and with the prediction results a warning message is sent to the sender if the text data in the social media contains bullying kind of messages. Also, a block notification is sent to the receiver asking him to block the text from the sender.

Index Terms— Cyberbullying, Classification algorithms, Performance measures, Notification.

Introduction

In recent days social media acts as a platform enabling users to communicate with one another and share the images, textual messages and videos. The rapid extension of internet and technology has made people increasingly depend on social media platforms in their daily lives [6]. Cyberbullying on social media has become a common issue in the modern digital age and creates a lot of negative impacts on people. Cyberbullying emerged as a pervasive issue on social media platforms and this dangerous form of online harassment cause deep emotional wounds with potential outcomes including mental health challenges, social exclusion, and even suicide. It has been observed that the victims on cyberbullying rate ranges from 10% to 40% [7]. The extensive impacts on victims emphasize the urgency to think up effective measures for detection and prevention of such bullying messages. Exploiting the capabilities of machine learning becomes instrumental in finding out the complex language patterns employed by cyberbullies, thereby enabling the creation of an automated model to promptly identify such harmful actions. This proposes a robust supervised machine learning framework explicitly designed to detect and prevent cyberbullying instances. Through the utilization of multiple classifiers, the models not only impart the ability to identify bullying behaviors but also empower the system to offer indispensable support to those affected. The versatility of machine learning extends beyond preventing acts of cyberbullying, as it can play a pivotal role in extending aid to victims.

Machine Learning is an area in computer science that trains the machines to learn the data patterns, current trends and make decisions using various algorithms. The advent of various machine learning as well as the natural language

processing algorithms makes ease in exploring characteristics of cyberbullying messages and classify them comparing with the preprocessed data.

Literature Survey

The authors in [1] have used both bag of words feature and TF-IDF feature to recognize various bullying words and passed those data to few of the learning algorithms which includes decision tree, support vector machine and random forest thereby observing that support vector machine model performs better for the twitter dataset. In [2], the authors have analyzed few classification algorithms in terms of performance measure and demonstrated that Logistic regression algorithm performs better with an accuracy of 90%. The authors in [4] conducted the study of classifying the data into bullying message or not using some of the supervised and few of the ensemble machine learning algorithms on the twitter dataset from kaggle. In [5], the authors have used some of the machine learning algorithms and observed that logistic regression performs better comparing the performance metrics.

Methodology

Proposed System

The major idea of this system is creating an effective cyberbullying detection model which uses learning and language processing algorithms (word tokenization, word lemmatization, stop words removal) in tracing out the intricate language patterns used by cyberbullies to prevent the harmful actions. These algorithms makes ease to explore the characteristics of cyberbullying messages and classify them comparing with the preprocessed data.

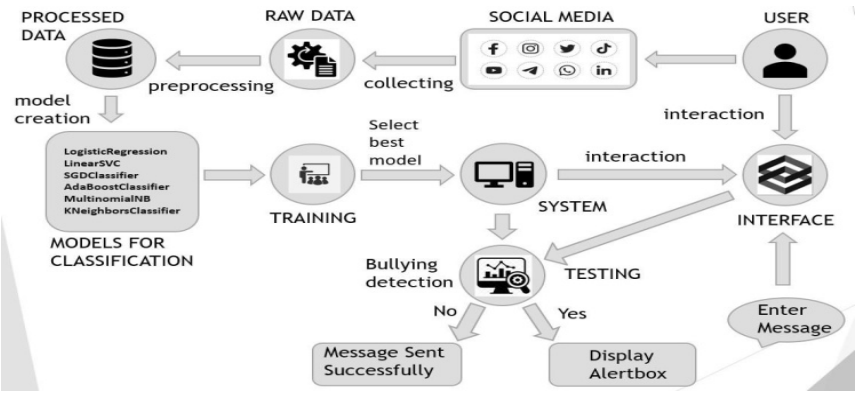


Figure 1: System Architecture

Data Preprocessing

Data Collection. It is the basic method of collecting the data and training machine learning models. The data which is trained data predicts frequent patterns thereby confining the facts about the past events. The models can be built using machine learning algorithms making use of the patterns which predict the input text as a bullying message or not. The dataset is generated using different comments found across the various social media platforms and concatenating them along with the already existing dataset from kaggle.

Data Cleaning. Data preprocessing is the process of cleaning the data and from the dataset it is observed that the data contains text with special characters and improper punctuation and grammar. As a part of the first step, the data is processed to eliminate the unnecessary characters, symbols and special characters from the data [1]. The next step is to convert all the text data to lowercase letters to ensure consistent comparisons. In the third step tokenization is performed which splits the text into individual words or tokens for further analysis while the fourth step is to remove the stopwords which helps in eliminating the common words. The last step called lemmatization is used to reduce words into their root form and this process is done to avoid duplication of similar words.

Machine Learning Algorithms Implementation

Stochastic Gradient Descent algorithm

This algorithm is mainly used to solve large-scale classification tasks where the parameters used in the model are iteratively updated to minimize the loss function with a randomly selected subset of training data. This makes the algorithm efficient and suitable for handling massively large datasets and the iterative nature of the algorithm makes it to converge to an approximate solution.

Logistic Regression.

This algorithm takes the input features and estimates the probability of binary values. A logistic function transforms features into the values ranging between 0 and 1 thereby signifying the likelihood of one class over the other.

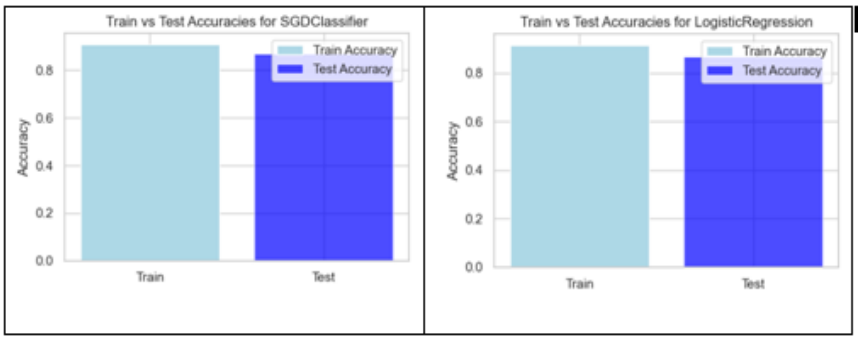


Figure 2: Performance Evaluation using SGD Classifier and Logistic Regression

Multinomial Naive Bayes Classifier.

This classifier is a probabilistic method that performs classification tasks assuming the features are conditionally independent given the class label and estimates likelihood of each feature's occurrence in different classes using a multinomial distribution.

Random Forest.

The ensemble learning method which integrates strength of multiple decision trees thereby increasing the prediction values and ease the over fitting risks. With bootstrap aggregation each decision tree is nurtured with a distinctive subset of training data imparting them unique perspectives. These trees collaboratively harmonize their predictions, culminating in a final output achieved through averaging or democratic voting.

Figure 3: Performance Evaluation using Multinomial NB and Random Forest

Results and Discussions

Results

The table 1 mentioned below shows the performance measures [8] of various classification algorithms.

Table 1: Performance Evaluation Results

Sl.No	Machine Learning Algorithms	Performance Measures			
		Accuracy	Precision	Recall	F1 Score
1	Stochastic Gradient Descent	87.13	93.24	84.46	88.63
2	Logistic Regression	86.70	92.41	84.55	88.31
3	Support Vector Classifier	85.40	89.700	85.20	87.40
4	Multinomial Naive Bayes	82.90	84.46	87.28	85.85
5	K-Nearest Neighbor	77.69	96.07	65.10	77.61
6	Decision Tree	76.33	98.94	60.81	75.32
7	Random Forest	59.69	59.59	99.81	74.63

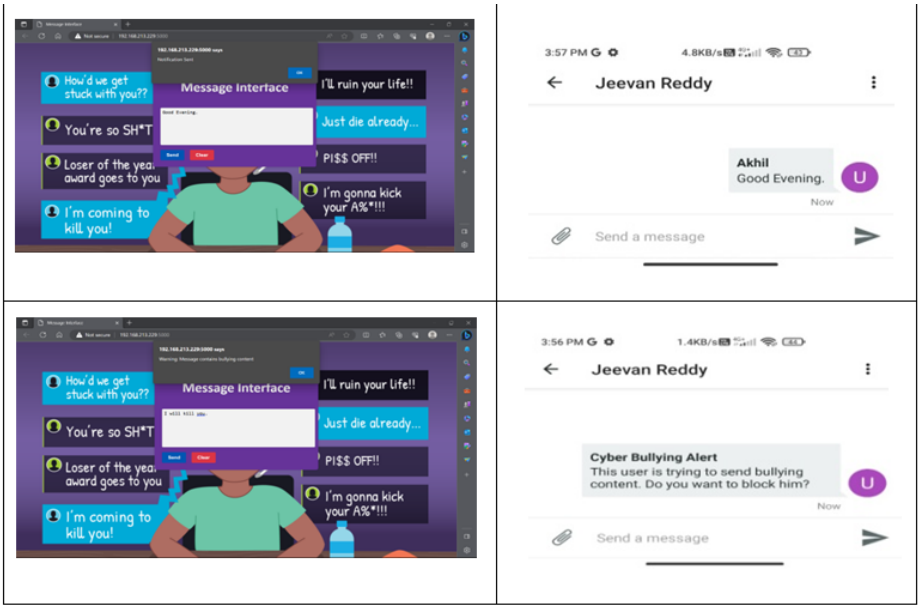


Figure 4: Notification message indicating the bullying text.

Conclusion

The proposed study focused on implementing the various machine learning algorithms that suggests in creating a safe digital landscape. A robust system is developed that effectively discriminates the instances of cyberbullying within the expansive realm of social media interactions. From the study it is observed that stochastic gradient descent classifier performs better and with the prediction made by the model a notification is sent to the sender if the data contains bullying messages and a block notification is sent to the receiver indicating the characteristics of bullying messages and asking him to block his number.

References

1. M Auddin, L Islam, S Sharmin: Cyberbullying Detection on Social Networks Using Machine Learning Approaches. (2021).

2. Trana R.E., Gomez C.E., Adler R.F: Fighting Cyberbullying: An Analysis of Algorithms. Advances in Intelligent Systems and Computing, vol 1, 213. (2021).

3. R. R. Dalvi, S. Baliram Chavan and A. Halbe, Detecting A Twitter Cyberbullying Using Machine Learning, ICICCS. (2020).

4. S Suleiman, P Taneja, A Nainwal: Cyberbullying detection on twitter using machine learning: A review, International journal of Innovative Science and Research Technology, Vol 7, No. 6, (2022).

5. A Muneer and S M Fati: A comparative analysis of machine learning techniques for cyberbullying detection on twitter.

6. C Fuchs: Social media: A critical introduction. Sage. (2017).

7. S Kemp: Digital 2019: Global Digital overview - global digital inights. (2019).

8. S. K. Shanmugam, S J. Devi: A study on performance of classification models for Covid-19 Datasets. Volume 12 (Number 10): 1123-1127 (2021).