iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** IX **Month of publication:** September 2021

DOI: https://doi.org/10.22214/ijraset.2021.38264

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue IX Sep 2021- Available at www.ijraset.com

Cyberbullying Detection System Using Machine Learning

Ms. Shama Kabeer
P.G Student, Department of Computer Science & Engineering, Thejus Engineering College, Kerala, India

Abstract: Cyberbullying is an online form of harassment. By posting, commenting, sending, or distributing personal, derogatory, false, or nasty stuff about others that can shame or humiliate them, this conduct is done with the goal of harming others. Once such content is published on the internet, it remains accessible indefinitely. This activity is considered unlawful, and it is more widespread among children and teenagers. Cyberbullying is an online epidemic that has the potential to result in devastating outcomes such as violence and suicide, and so must be dealt with swiftly and properly. To detect bullying behavior in textual messages, a real-time cyberbullying detection system based on machine learning—Naïve Bayes Algorithm is presented. The model was created to determine whether a tweet was bullying or non-bullying in nature. Also, to assist victims in dealing with bullying difficulties without their identities being revealed.
Keywords: Machine Learning, Cyberbullying, Naïve Bayes, Cybercrimes, Cyberbullying Detection

I. INTRODUCTION

With the advancement of technology, internet has become a safe and secure space for communication. But as the social media lifestyle surpasses the physical barrier of interaction and spare the space for inappropriate interaction with unknown people, it is highly prone to cybercrimes. Cyberbullying is bullying that occurs online and can occur on a variety of sites where people read, participate in, or contribute content. Bullying or harassment is defined as a pattern of behaviour with the goal to cause harm to others. Derogatory, threatening, or harassing communications, photos, audios, and videos are examples of cyberbullying. Once such stuff is published, it exists indefinitely in cyberspace. Cyberbullying allows a bully to humiliate and damage a victim anonymously in online forums due to the simplicity with which such stuff may be posted. Furthermore, victims and onlookers are afraid of being punished or made a social outcast if they disclose instances. Bullying is more prevalent among children and teenagers. Cyberbullying has a severe effect on this group, and victims often have low self-esteem as a result. Bullying can have a variety of negative consequences, including poor effects on mental and physical health, sadness and anxiety, and even suicidal thoughts. Victims of cyberbullying may miss or even drop out of school as a result of such behaviour. As a result, cyberbullying has become an epidemic that must be addressed immediately and efficiently.

Facebook and Twitter, for example, provide tools and tactics that might assist users in reporting bullying and so promote a safe online experience. These include options for determining the intended audience, blocking specific users, and reporting and deleting people who engage in inappropriate behaviour. These strategies, while crucial, are reactionary in nature, meaning they occur after a person has already been abused. Many people may have read the offending post by the time a person reports it and the authority takes corrective action, and therefore the harmful impacts may have occurred. As a result, we require a method for quickly detecting cyberbullying conduct. People of all ages nowadays utilise the Internet, blogs, forums, and social networking sites extensively. Bullies have moved from the outer world to the Internet as the use of several digital communication media has grown. Bullying is defined as a pattern of international behaviour by a group or an individual that makes it impossible for the victims to defend themselves. Cyberbullying is a new type of bullying that involves using the Internet, smartphones, and other devices to email or post text or photographs with the goal of injuring or embarrassing the victims. Cyberbullying, on the other hand, can be viewed as a technique of using social media in an illegal and immoral manner.

Multiple machine learning, deep learning, and their integration strategies have been identified in the literature. To develop the predic tion and classification model, the study had to reduce the number of algorithmsto just the Naïve Bayes algorithm.For the implement ation, many approaches such as CNN, SVM, Fuzzy Fingerprints, SelfAttention Model, Data Mining, Text Mining, and Neural Netw orks are used. However, the majority of the approaches failed during the classification and detection processes. Some of the strategi es were also unable to be used due to their high implementation costs. As a result, there is a larger need for a more accurate model th at can take into account all of these constraints and overcome them in order to provide an adequate solution to the problem of detecti ng social media bullying.

II. RELATED WORK

The detection of cyberbullying is helpful in the development of technology to protect people on social media sites. Because earlier work on cyberbullying detection is important to our topic, we detail it here. Many scholars have advocated employing various approaches and technology to detect bullying language.

B. Kansara et al. proposed a novel system for detecting abusive content through online interactions. To analyse the adult photos, they used the Bag of Visual Words (BoVW) concept in conjunction with the SVM classifier. To classify the abusive text messages, they combined Bag of Words (BoW) with the Nave Bayes classifier. The results of the experiments revealed that their proposed strategy performed effectively not just with text messages but also with image analysis. They had the highest memory rate of the bullying content, with a 66 percent classification rate.

D. Viktor et al. proposed a strategy for detecting aggression text by mixing shallow and deep learning approaches. They experimented using data from Facebook and Twitter. They employed a support vector machine (SVM) classifier and a bidirectional long short term memory network (BiLSTM). They received the highest F1 score, 0.616 for Facebook data and 0.565 for Twitter data.

H. Qianjia et al. analysed the use of convolutional neural networks (CNN) combined with natural language processing technologies to detect bullying in real time was investigated. They also gave a way for individuals to provide early feedback on the word choice before it was shared on social media. In terms of F1 score, they obtained 0.517 for the Support Vector Machine (SVM) model, 0.585 for the combination of SVM and Linguistic Inquire and Word Count (LIWC) model, 0.523 for the CNN model, and 0.597 for the combination of CNN and LIWC model.

Zhao et al. proposed a framework for detecting cyberbullying using word embedding, which creates a list of pre-defined insulting phrases and assigns weights to obtain bullying characteristics, and they utilized SVM as their main classifier, which had a recall of 79.4 percent.

III. THE PROPOSED METHOD

The step-by-step processes for detecting cyberbullying in textual data – from preprocessing to final analysis of the framework – are presented in this chapter. The Naïve Bayes method is used to anticipate the existence of bullying/non-bullying in tweets/comments. It is a supervised learning classification technique that is based on the premise of Bayes' theorem and can be used for rapid classifications. The key benefit of employing the Nave Bayes method is that it is a simple and effective classification technique that aids in the development of a fast machine learning model capable of making quick predictions.

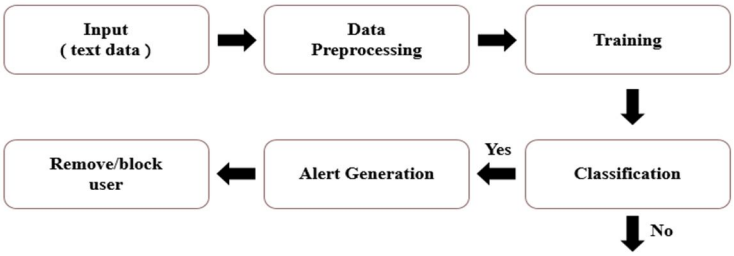


Fig. 1 The block diagram of the proposed cyberbullying detection system

The project's proposed architecture is depicted in fig.1 above. It is made up of several blocks, including input, data preparation, training, classification, alarm production, and the final block, which receives the victim's response. The user's input, in the form of tweets/comments in text format, is taken first. It is then submitted for BoW conversion, in which the stop words are eliminated, the grammar is ignored, but the sense of the words is preserved. After conversion, the output is presented for categorization; if it is classified as a bullying message, the victim is notified; otherwise, it is not. The attacker's account is then deleted/blocked in accordance with the victim's wishes. In this case, 80% of the data is used for training and 20% for testing.

A. Data Preprocessing

This module's primary goal is to process the tweets collected from the dataset. The data retrieved from the database might not be in its purest or most standard form. As a result, we must execute several preprocessing processes on the dataset before letting the model learn on its own. Dataset collecting, data analysis, data cleaning, tokenization, and data splitting are some of the procedures that must be completed under this module.

1) *Dataset Description:* For this approach, a cyberbullying twitter dataset was collected from kaggle. This dataset was originally created to identify and classify toxic online comments. It contains 1064 samples of comments. From this dataset 80% of the data was used for the training process and 20% was used for the testing process.
2) *Data Analysis:* In this step the tweets that are collected from the dataset are analyzed. All the dimensions are analyzed and identified from the dataset. This step is very essential in order to build up an understanding of various relationships among the terms mentioned with their dimensions in the dataset.
3) *Data Cleaning:* The data in the dataset may be impure because it contains a lot of undesirable or unnecessary terms and attributes. As a result, these meaningless phrases are referred to as noise. As a result, removing certain terms before feeding the data into the model for training is critical. All unnecessary characteristics are deleted and discarded, leaving only the essential features to be extracted from the dataset. This is known as feature extraction.
4) *Data Tokenization:* This step consists of representing each word as a vector of real values known as word embedding vectors. This allows us to compute distributed representations of words in the form of continuous vectors. Word2Vec can be used for explicitly encoding many semantic relationships in addition to linguistic regularities and patterns into the new embedding space.
5) *Data Splitting:* Finally, the dataset must be separated for training and testing the model. A larger percentage , i.e., 80 % of the dataset is chosen for training purposes, while the remaining , 20% is used to test the data.

B. Model Building

Under this module, a model that can predict and detect whether a tweet is bullying or not has to be constructed. This model can be developed using the Naïve Bayes Algorithm, which is a supervised classification algorithm.

C. Training and Testing

Following the creation of the cyberbullying detection model, the model may be trained using the training data from the training dataset. Following that, the model's validity must be determined by combining the validation data with the model's training to establish the training accuracy. Finally, some test data from the test dataset can be used to test the completely formed model.

D. Twitter API Workflow

It is feasible to integrate Twitter with the completely constructed model utilising various twitter APIs supplied by twitter as an extension to the fully built model. Using Twitter's APIs, we can gather real-time data or tweets. The programme uses the collected tweets as inputs to analyse and detect the presence of bullying. The trained model can then quickly analyse the tweets and determine whether or not they are bullying, and take appropriate response based on the victim's needs.

IV. RESULT AND ANALYSIS

The system is built to detect cyberbullying behavior in textual messages, real-time detection system, by using machine learning approach and also to analyze the performance of the system. The model was evaluated using various metrics which are Accuracy, Precision, Recall and F1-score to assess the system performance. All these metrics are calculated during both the training and testing phase and their minimum and maximum values are collected. The time taken for training and prediction is also evaluated.
Accuracy is the most widely and frequently used matric for evaluating the performance of a classifier. It Measures the accurately predicted samples from the total number of samples. It shows the fraction of correctly predicted samples out of the total number of sample inputs. Here, this metric measures the number of tweets correctly classified. Mathematically, it can be represented as,

Accuracy = (TP + TN) / T

In machine learning precision can be defined as the percentage of instances from the total available instances that were classified correctly or are relevant. It measures the proportion of positive identifications that are correctly classified. Here, precision measures the number of tweets classified by the algorithm as bullying and are actually proved to be bullying tweets. Mathematically it can be represented using the following formula,

Precision = TP / (TP + FP)

The Recall measure can be defined as the percentage of actual positives that were correctly classified. It is the proportion of actually relevant instances retrieved among the relevant data. It can also be termed as Sensitivity in machine learning. Here, this metric measure how many bullying tweets, out of all the available ones are actually detected by the algorithm. Mathematically it can be represented as,

Recall = TP / (TP + FN)

F1-score is another evaluation metric which is used to maintain a balance between the above two metrices, Precision and Recall. It is used to find the accuracy of any test based on the average of the precision and recall scores by considering its values. Here, this metric is computed using the harmonic mean of precision and recall. It can be represented using the values of precision and recall as,

F = 2 * (Precision * Recall) / (Precision + Recall)

Training	**Min**	**Max**
Accuracy	**0.89**	**0.99**
Precision	**0.92**	**0.99**
Recall	**0.88**	**0.99**
F1-score	**0.92**	**0.99**

Table1.Performace metrics during training.

Testing	Min	Max
Accuracy	0.85	0.92
Precision	0.89	0.97
Recall	0.88	0.94
F1-score	0.89	0.94

Table2.Performace metrics during testing.

Table1 and Table2 show the minimum and maximum values of all the metrics during training and testing period.

V. CONCLUSION

In this paper, a cyberbullying detection system that detects whether a tweet/comment is bullying or non-bullying is proposed. This project is designed in such a way that it also helps the victims to address the issues of bullying without their identity getting exposed. Cyberbullying is the bullying that takes place in the digital world, it is a repeated behavior and an intended to harm others. Once, such a content is posted online, they live perpetually in the cyber world. This act is most common among kids and youngsters, which can end tragically in suicides and violence. Therefore, cyberbullying in an epidemic that needs to be controlled quickly and effectively. The algorithm is used for the implementation of this project is Naïve Bayes Classification algorithm, which is a supervised classification algorithm that can be used for quick predictions in texts. In the result of this project, initially the input given as text will be classified as either bullying or non-bullying and then the victim can report/block the attacker as per their requirement. There are number of further directions that can be investigated beginning from this project.

VI. FUTURISTIC SCOPE

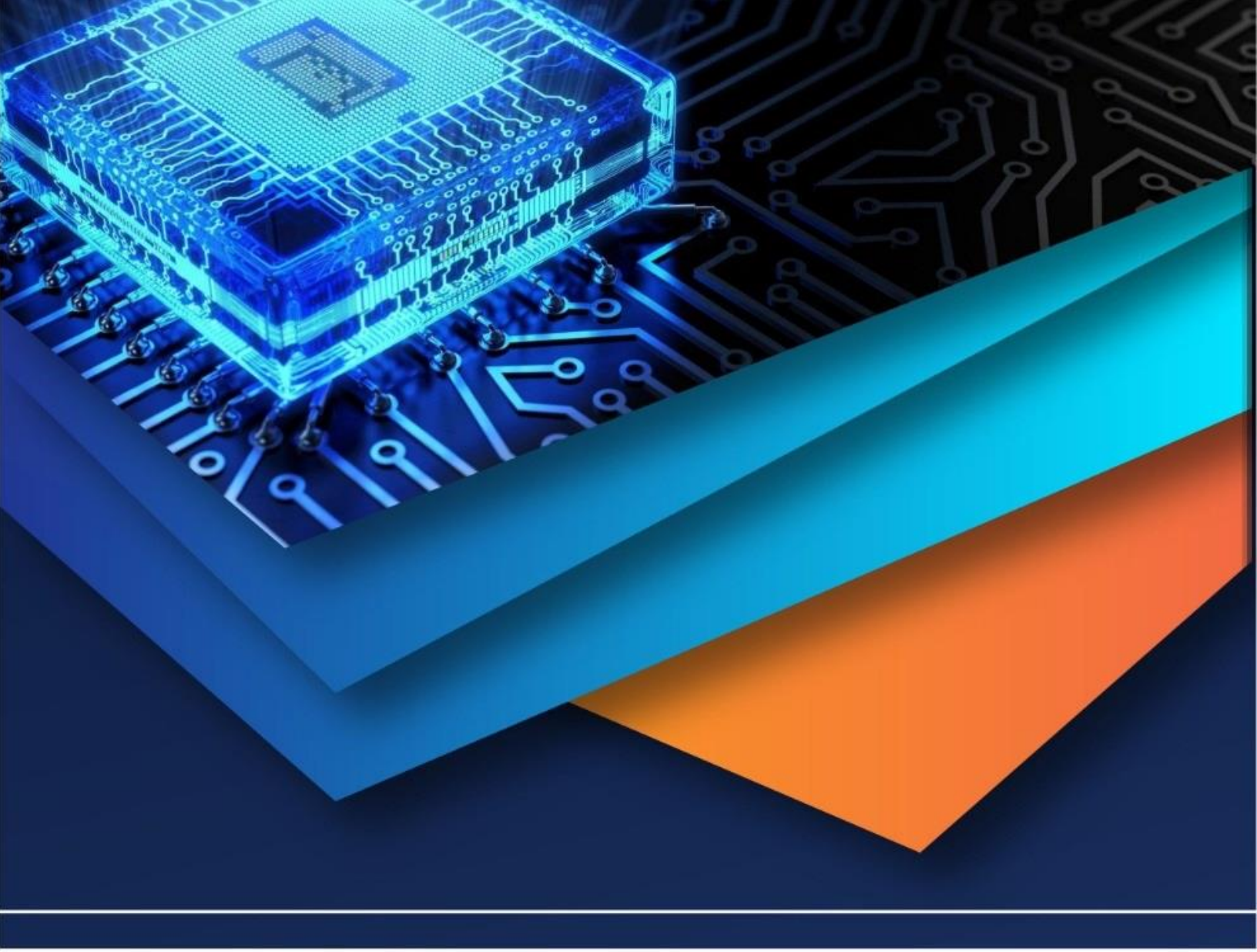
There are number of further directions that can be investigated beginning from this project. The first one is to include other social media sites along with Twitter to extend the usage of the system. Second, the dataset only consider English language, so it can be made multilingual. Third, this approach can be validated on very large dataset. Fourth, by the incorporating approaches like Natural Language Processing (NLP), the sentiment and sarcasm associated with each input can be detected. Finally, as of now this project only concentrates on textual data, this can be extended to detect malicious activities in images, videos and gifs.

VII. ACKNOWLEDGEMENT

I would like to express my gratitude to my project coordinator, Mr. Valanto Alappatt, Assistant Professor of Computer Science and Engineering Department and my project guide Ms. Sindhu S, Assistant Professor of Computer Science and Engineering Department who guided me throughout this project. I would also like to show my deep appreciation to our HOD Mrs.Vinitha A.V who helped me finalize my project. I am extremely thankful to Dr. K Vijayakumar, Principal of Thejus Engineering College, Vellarakkad, who gave me the golden opportunity to do this project. I wish to acknowledge the help provided by the technical and support staff in the Computer Science department of the Thejus Engineering College. Finally, I want to express my gratitude to my parents and numerous friends who have supported and loved me throughout this long process.

REFERENCES

[1] Shiza Andleeb, Dr. Rashid Ahmed, Zaheer Ahmed, Maira Kanwal, "Identification and Classification of Cybercrimes using Text Mining Technique." 2019 International Conference on Frontiers of Information Technology(FIT).
[2] Saloni Mahesh Kargutkar, Prof. Vidya Chitre, "A Study of Cyberbullying Detection Using Machine Learning Techniques." Proceedings of the Fourth International Conference on Computing Methodologies and Communication(ICCMC 2020).
[3] Yuzana Win, "Classification using Support Vector Machine to Detect Cyberbullying in Social Media for Myanmar Language." 2019 IEEE International Conference on Consumer Electronics- Asia (ICCE-Asia).
[4] Vijay Banerjee Jui Telavane Pooja Gaikwad Pallavi Vartak , "Detection of Cyberbullying Using Deep Neural Network." 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS).
[5] Sourabh Parime, Vaibhav Suri, "Cyberbullying Detection and Prevention: Data Mining and Psychological Perspective." 2018 International Conference on Circuit, Power and Computing Technologies [ICCPCT].
[6] Hugo Rosa, Joao P. Carvalho, Pavel Calado, Bruno Martins, Ricardo Ribeiro, Luisa Coheur, "Using Fuzzy Fingerprints for Cyberbullying Detection in Social Networks." 2018 IEEE International Conference on Fuzzy Systems (FUZZ).
[7] Yash R Mahadik, "Sarcasm Detection using Support Vector Machine.", Volume 8 Issue XI Nov 2020 International Journal for Research in Applied Science & Engineering Technology (IJRASET).
[8] Mr. Shivraj Sunil Marathe, Prof. Kavita P. Shirsat "Contextual Features Based Naïve Bayes Classifier for Cyberbullying Detection on YouTube" International Journal of Scientific & Engineering Research, Volume 6, Issue 11, November-2015.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY