INTERNATIONAL JOURNAL OF ADVANCED INNOVATIVE TECHNOLOGY IN ENGINEERING

Published by Global Advanced Research Publication House Journal

Home page: www.ijaite.co.in

Cyberbullying Detection System using Machine Learning in English

[1]Gulnaz Sheikh, [2]Himanshi Aglawe, [3]Jagdeep Singh Arora, [4]Mohammad Tauheed Raza Patel, [5]Ossama Ansari, [6]Ruchik Ganvir, [7]Prof. Sadia Patka

[1,2,3,4,5,6,7]Department of Computer Science & Engineering, Anjuman College of Engineering and Technology, RTMNU Nagpur, Maharashtra, India

[1]gulnazsheikh09t@gmail.com, [2]himanshiaglawe9858@gmail.com, [3]jagdeepsingharora19@gmail.com, [4]tauheed1patel@gmail.com, [5]ansariossama@gmail.com, [6]ganvir.ruchik1404@gmail.com, [7]spatka@anjumanengg.edu.in

Article History

Received on: 01 March 2022

Revised on: 14 March 2022

Accepted on: 31 March 2022

Keywords: Natural Language Processor (NLP), Natural Language Toolkit (NLTK)

e-ISSN: 2455-6491

DOI: 0.5281/zenodo.6674769

Production and hosted by
www.garph.org

ABSTRACT

Nowadays, users are demanding security while using social media. The increased use of social media by teenagers has led to cyberbullying becoming a major issue. The purpose of the study is to explore some of the different variables that influence people to become cyberbullies. The abstract idea of the paper deals with the idea of detecting harsh comments related to cyberbullying done on social media using Machine Learning algorithms and an NLP. For performing NLP, we used the NLTK library in our detection system. The various references are used for understanding the paper-related algorithms and the project. In this paper, we have used the Naive Bayes algorithm for classification and regression.

1. INTRODUCTION

In the world of Computers, there is a vast usage of social media and nowadays many people are using social media. Hence this Cyberbullying has been a major cause of worry for the amount of serious impact it has on people. Although social media is a secure place for communication also sometimes it is prone to cyberbullying. It is found to be more dangerous than traditional bullying because the humiliation is visible to an unlimited online audience. Since the physical appearance of the victim is not required it can go on nonstop. Many networking sites don't even need a real name to be registered as a user making the bullies braver. The victims who have undergone bullying lose their self-confidence and become antisocial and this has a bad effect on their mental health as well. Social media is being extensively used today. This has led to a form of bullying that is Cyberbullying. Bullies use various network sites to attack victims with offensive comments and posts.

While humans also provide critical serviceable oversight and brilliant insights into today's infrastructure, machine learning and artificial intelligence are rapidly gaining traction in almost all domains of today's systems, be it on-premises or on the cloud. [2]

Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values. Machine learning algorithms build a model based on sample data, known as training data, to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, speech recognition, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

Classical machine learning is often categorized by how an algorithm learns to become more accurate in its predictions. There are four basic approaches: Supervised learning, Unsupervised learning, Semi-supervised learning, and Reinforcement learning. The type of algorithm data scientists chooses to use depends on what type of data type want to predict.

- In this context, we suggest a cyberbullying detection model based on machine learning that can detect whether a text relates to cyberbullying or not. [1]
- Cyberbullying Detection implements our coded, machine learning algorithms, in finding a negative comment from the message it receives from a user. [1]
- Our cyberbullying detection system is based on 3 modules i.e., data preprocessing, feature extraction, training, and testing.

2. LITERATURE REVIEW

Many approaches propose systems that can detect cyberbullying automatically with high accuracy. In 2021, Rupesh Kumar, Shreyas Parakh, C.N.S. Vinoth Kumar proposed the Detection of Cyberbullying using machine learning. The experimental outcomes of the proposed method are elaborated on the cyberbullying dataset from Twitter. The Twitter cyberbullying dataset had been compiled and labeled. In general, this dataset includes a large number of Twitter communication posts. The machine learning algorithms which are used are naïve Bayes and Support vector machines. In the system, Naïve Bayes and SVM are used because they are among the best performing classifiers available. After performing multiple studies using various n-gram language models. At the time of estimation of the model generated by the classifiers, it paid special attention to 2-gram, 3-gram, and 4-gram. [2]

In 2021, Md Manowarul Islam, Md. Ashraf Uddin, Linta Islam, and Arnisha Akter proposed the system named Cyberbullying Detection on social networks using Machine Learning Approaches. In the proposed system, the Machine learning algorithms which are used are Decision tree, Naïve Bayes, Random Forest, and Support vector machine algorithms. Dataset was built from the user comments on different Facebook posts. They compared the various parameters of the machine learning algorithms based on the two important features vectors BOW and TF-IDF. After the precision and accuracy results, it was clear that SVM outperforms the other algorithm. The result also indicated that TF-IDF provides better accuracy than BOW features. [3]

In 2020, Vimala, and Balakrishnan presented an automatic cyberbullying detection taking Twitter users' psychological features into account. The three main stages discussed in improving cyberbullying detection are Twitter data collection, feature extractions, and cyberbullying detection and classification. The annotated dataset contained 9484 tweets, out of which 4.5% of users are labeled as bullies, 31.8% as spammers, 3.4% as aggressors, and 60.3% as normal. However, the final dataset contained 5453 tweets as a result of the pre-processing step which included removing non-English tweets, profiles containing no data, and special characters. The features extracted were text features, user features, and network features. The model was executed using WEKA 3.8 with 10-fold cross-validation. Since Naïve Bayes performed poorly during preliminary experimental analysis it was eliminated while Random Forest and J48 continued to perform well. The classifiers were trained using manually annotated data. [4]

This dataset is built from the user comments on different Facebook posts. We compare the various parameters of the machine learning algorithms based on the two important features vectors BoW and TF-IDF. The results also indicate that TF-IDF provides better accuracy than the BOW feature. This is because rather than taking almost all word into vectors, TF-IDF takes the most frequent words and maintain better performances.

In 2020, Sudhanshu Baliram Chauhan proposed an approach to detect cyberbullying on Twitter. The required dataset was collected from sources like GitHub, and Kaggle. Initially, the data is pre-processed and features are extracted using a TFIDF vectorized algorithm. These tweets are then passed through the naive Bayes and SVM model and are classified accordingly. When a tweet is categorized as bullying, ten other tweets from that user's account will be fetched and passed through naive Bayes and SVM classifiers again. If the overall probability of that user's tweets lies above 0.5 then it will be considered a bullied tweet. Based on the accuracy score and the results it was evident that the SVM model outperformed the naive Bayes with an accuracy score of 71.25%. [5]

In 2019, John Hani et al. [6] presented a supervised learning approach to detect cyberbullying. As a part of the pre-processing step, data is cleaned by removing the noise and unnecessary text. This is performed using tokenization, lowering text, and stop words along with encoding cleaning and word correction. The second step is the feature extraction step which is done using TF IDF and sentiment analysis technique including NGrams for considering different combinations of the words like 2-Gram, 3-Gram, and 4-Gram. The cyberbullying dataset from Kaggle is split into ratios (0.8, 0.2) for train and test. SVM and Neural networks are used as classifiers that run on a different n-gram language model. Accuracy, recall and precision, and f-score are the performance measures. It is found that Neural Network performed better than the SVM classifier. Neural Network achieved an average f-score of 91.9% and SVM achieved an average f-score of 89.8%.

In 2018, Monirah Abdullah Al-Ajlanet and Mourad Ykhlef [7] proposed a novel algorithm CNN-CB which is based on a convolutional neural organization and adapts the idea of word embedding. The architecture comprises four layers - Embedding, Convolution Layer, Max Pooling Layer, and Dense Layer. The first layer, word embedding, creates a vector space of vocabulary which is the input to the subsequent layer, the convolutional layer, which compresses the input vector without losing significant features. The third layer, the Max pooling layer, takes the output of the second layer as its input and finds the maximum value of the chosen region to save just significant highlights. The last layer, the Dense layer, does the classification. This gave a precision of 95%.

In 2018, Monirah A. Al-Ajlan et al., proposed optimized Twitter cyberbullying detection based on deep learning (OCDD) which does not extract features from tweets instead, it represents a tweet as a set of word vectors that are fed to a convolutional neural network (CNN)for classification. Hence the feature extraction and selection phases are eliminated in this approach. To represent the semantics between words, word embedding is used and is generated using the (GloVe) technique. CNN uses a lot of parameters and to optimize these values, a metaheuristic optimization algorithm is used to find optimal or near-optimal values that will be used for classification. CNN showed great results. [8]

In 2017, Yee Jang Foong and Mourad Oussalah [9] presented an automated cyberbullying detection that uses natural language processing techniques, text mining, and machine learning. For dataset ASKfm, a social media platform where users can anonymously ask questions and view a sample of a user's profile is used. As a part of the pre-processing procedure web links and unknown characters are removed, incorrect wordings in case any are corrected, and also lexicons are replaced with equivalent textual expressions. A combination of features has been used which includes TF-IDF, Unusual capitalization count, LIWC, and Dependency parser. The data set is split into a 70% training set and 30% testing set. SVM was used as a classifier which was trained with a linear kernel on the training data. To label the training posts Amazon Mechanical Turk Service was used. The combination of features mentioned above yielded the highest performance in terms of accuracy, precision, recall, F1, and F2 scores.

3. Proposed Methodology

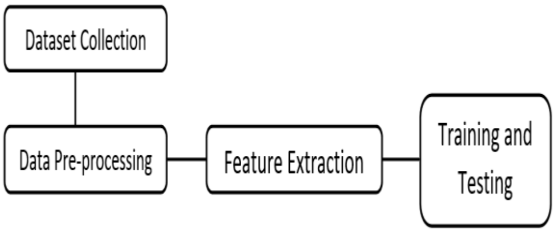

Figure 1: Steps of Proposed System

In this research on cyberbullying, detection is used for detecting abusive language in social media texts. To emulate social media platforms, a text detection system is developed in python using machine learning. Abusive language is detected in backend processing with the Naïve Bayes algorithm using the GaussianNB library (one of the Naive Bayes classification algorithms) from sklearn. naive Bayes. The system will detect the accuracy using our proposed algorithm. If the system recognizes some abusive or harsh comments in the given dataset, then our system will detect them using our proposed algorithm.

A. *Dataset Collection*

The first step is dataset preparation. We used the harassment dataset from the Kaggle site. The dataset accommodates 8799 rows with 4 columns. These statements are congregated from the kaggle_parsed_dataset (harassment dataset).

B. *Data Preprocessing*

In data pre-processing, we cleaned the data using different three methods one of the methods we used is the cleanest () method with the help of regular expression. We imported the library and used it. After that, we also replaced most of the unwanted signs with the space. The duplicated data were then removed using the duplicated method for the same. After data Pre-processing, we are left with 7498 rows in the dataset.

C. *Feature Extraction*

For feature extraction, we have used the vectorization technique. Vectorization is jargon for a classic approach of converting input data from its raw format (i.e., text) into vectors of real numbers which is the format that ML models support. This approach has been there ever since computers were first built, it has worked wonderfully across various domains, and it's now used in NLP. In Machine Learning, vectorization is a step-in feature extraction. The idea is to get some distinct features out of the text for the model to train on, by converting text to numerical vectors.

D. *Training and Testing*

For training and testing, we split the dataset using python libraries, after splitting the dataset we tested the data using a confusion matrix. We got to know how much data is taken by the machine for being predicted true or false. The testing is performed using Naïve Bayes is used in the project. After calculation, the calculated accuracy of the system is coming 60%. It can be further increased using different algorithms used for classification. The testing also contained the confusion matrix giving us an idea about how many values or taken data are being predicted true and how much data is being predicted false by the machine. For the confusion matrix, we used the confusion_matrix library from sklearn metrics. The parameters required for the calculation of the confusion matrix were the data we gave for testing and the predicted data we tested using the Naïve Bayes algorithm. Using the confusion matrix, we got the result as.

4. RESULT

On training and testing, we got to know that 2888 values are being predicted as true positive values. In our Cyberbullying Detection System, we got 60% accuracy using Naïve Bayes Algorithm.

2888 (True Positive)	1499 (False Positive)
1171 (True Negative)	1191 (False Negative)

CONCLUSION AND FUTURE SCOPE

The literature survey done in this paper provided insight into the detection of cyberbullying are mentioned which includes data collection, data preprocessing, feature extraction, and Training & testing. The proposed algorithm for the detection of cyberbullying by Naïve Bayes is successfully implemented. Our work can improve cyberbullying identification and help people use social media safely by achieving this level of accuracy. The proposed algorithm can prove to be highly useful in the real-time detection of cyberbullying and prevent emotional stress on victims.

As for future work, we would like to implement the proposed approach to detect cyberbullying in different languages as social media is vast and is not restricted to a single language. We can look for patterns in behavior on the social media platform instead of a single post. By identifying patterns, we can alert them based on the user's behavior. For future use, we can implement our proposed system for Audio detection and Image detection.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

FUNDING SUPPORT

The author declares that they have no funding support for this study.

REFERENCES

[1] Chanhee Shin, Giovanni Berrios: Social Media: A critical introduction. The Year of 2021

[2] Rupesh Kumar, Shreyas Parakh, C.N.S. Vinoth Kumar, "Detection of Cyberbullying using Machine learning", 2021

[3] Md Manowarul Islam, Md. Ashraf Uddin, Linta Islam, Arnisha Akter, "Cyberbullying Detection on social networks using Machine Learning Approaches", 2021

[4] Balakrishnan, Vimala & Khan, Shahzaih & Arabnia Hamid, “Improving cyberbullying detection using Twitter Users Psychological features and Machine Learning”, 2020

[5] R. R. Dalvi, S. Baliram Chavan, and A. Halbe, "Detecting A Twitter Cyberbullying Using Machine Learning", 2020

[6] John Hani, Mohamed Nashaat, Mostafa Ahmed, Zeyad Emad, Eslam Amer, and Ammar Mohammed, “Social Media Cyberbullying Detection using Machine Learning” International Journal of Advanced Computer Science and Applications(IJACSA), 10(5), 2019.

[7] Munirah Abdullah Al-Ajlan and Mourad Ykhlef, “Deep Learning Algorithm for Cyberbullying Detection”, IJACSA, 2018

[8] M. A. Al-Ajlan and M. Ykhlef, "Optimized Twitter Cyberbullying Detection based on Deep Learning", 2018

[9] Yee Jang Foong and Mourad Qussalah. Cyberbullying system Detection and Analysis, 2017.