Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

Social Media Cyberbullying Detection using Machine Learning in Bengali Language

Rounak Ghosh
B.Tech Student
Computer Science and Engineering Department
SRM Institute of Science and Technology
Kattankulathur, Tamil Nadu, India

Siddhartha Nowal
B.Tech Student
Computer Science and Engineering Department
SRM Institute of Science and Technology
Kattankulathur, Tamil Nadu, India

Dr. G. Manju
Associate Professor
Computer Science and Engineering Department
SRM Institute of Science and Technology
Kattankulathur, Tamil Nadu, India

***Abstract*-- Cyberbullying is the utilization of technology as a medium to menace somebody. Social networking sites give a prolific medium for menaces who utilize these sites to assault or harass helpless youthful grownups. Through Machine learning, we can distinguish language designs utilized by menaces and create rules to consequently recognize digital harassing content. Most of the works related to cyberbullying detection using machine learning have been proposed on languages such as English, Chinese and Arabic. Very few works have been done on regional Indian languages. In this paper, we have proposed a model that recognizes cyberbullying content in an uncommon or rather regional Indian language such as Bengali.**

Keywords—Cyberbullying, machine learning, random forest, passive aggressive classifiers

INTRODUCTION

Cyberbullying is totally different from conventional harassing, however it is as yet tormenting. The results and risks continue as before, if not extended in their seriousness and span. Despite the fact that it happens through online sites rather than face to face, cyberbullying should be viewed as appropriately.

At the appropriate time, cyberbullying comes in different various structures. It doesn't really mean hacking somebody's profile or presenting to be another person. It likewise incorporates posting negative remarks about someone or spreading bits of hearsay to criticize somebody.

Cyberbullying or Social Media Bullying incorporates activities and measures to control, annoy or stigmatize any individual. These horrible activities are solemnly harming and can influence anybody effectively and seriously. They basically happen via web-based media, public gatherings, and other online sites.

PURPOSE

The most recent decade has seen a flood of cyberbullying – this tormenting isn't simply restricted to English yet in addition it occurs in different languages. An enormous crowd is a fruitful ground for cyberbullies henceforth, it is vital to distinguish cyberbullying in different dialects. Thereby, the proposed model that we are going to develop will help in detecting cyberbullying contents in one of the outspoken and regional languages in India, i.e. Bengali. Very few works considering cyberbullying has been done in this language. We are going to use the same cyberbullying models that were used in previous works for English language. In spite of the fact that execution and performance may fluctuate because of semantic contrasts among English and non-English substance, to battle such issues, this model proposes the utilization of machine learning algorithms and the consideration of user data for detecting digital harassing on Bengali text.

LITERATURE SURVEY

In the paper [1], they arranged a dataset utilizing a java program created to separate Bangla Text discussions from online media stages essentially Facebook and Twitter, other than the discussions they likewise gathered client's segment information from Twitter utilizing Twitter Rest Api which was marked physically and sack of words approach was fused for the model. They made a model utilizing machine learning algorithms like SVM, KNN, Naive Bayes and J48 followed by evaluating the exposition of the different algorithms. They played out the examinations in two stages, in the main stage they prepared and tried model utilizing the content discussions in Bangla Text and in the second stage they included both the content based highlights and user's information. SVM beat the wide range of various calculations in the two stages.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

One of only a handful few papers we found that attempted to recognize harmful Bangla text is paper [2], utilized a root level calculation to identify oppressive content and furthermore proposed unigram string highlights to improve result. To discover, with what sort of highlight their proposed calculation performed better, they evaluated the experiment with various different string property namely, unigram, bigram, and trigram. The word importance in single sentence is not taken into consideration in the unigram features. In any case, words which were more harmful, it tends to be discovered utilizing this element.

In the paper [3], they have created a dataset of text conversions and comment obtained from the comment sections of public posts of some popular Facebook pages. They labelled the dataset into two categories namely bully and non bully, the bully category is further divided into four sub-categories as sexual, troll, religious, and threat. They also added gender classification to the dataset mentioning who wrote the comment and to whom it was directed. Other user specific features like occupation were added to the comments in the dataset and lastly the number of reactions on the particular comment. For each column in the dataset they presented an analysis to provide further insights about the dataset. The dissemination and number of information in every classification give a decent wellspring of learning a decent machine learning model to recognize distinctive sort of cyberbullying in Bangla language.

With the help of Machine Learning it can be useful to recognize language examples of domineering jerks; furthermore it can accordingly develop a model to identify digital tormenting activities. In this manner, the fundamental commitment of paper [4] was to put forward a regulated Machine learning model for distinguishing along with forestalling digital harassing in English language. The model was evaluated on a dataset containing digital harassment contents from kaggle. The consequences of SVM and Neural Network classifiers were thought about on both TFIDF as well as sentiment analysis extraction methods. It was established that Neural Network performed better compared to that of SVM classifier. Besides, this work was contrasted with another connected work that utilized the equivalent dataset, tracking down that Neural Network outflanked their classifiers regarding precision and f-score.

In the paper [5], they proposed a model that provides a multilingual cyberbullying detection approach in different Indian languages mainly Hindi and Marathi, they created their own API and scrapper to collect the dataset from various social media platforms, newspaper reviews and tour reviews. After manually labelling the dataset and removing all redundancies, stop words and special characters they fed the dataset into the model. Machine Learning languages like Logistic Regression, Stochastic Gradient Descent and Multinomial Naive Bayes were used for the model. Using the bag of words approach which neglects the grammar as well as the sequence or occurrence of the words but keep hold of the frequency of the words, they classified the dataset. To measure the accuracy they calculated the f1 scores of each algorithm and found that LR outperformed all the other ML algorithms with the least error rate

In paper [6], they described an overview on multilingual cyberbullying recognition. After going through a lot of research, they founded that most of the work detecting cyberbullying has been done mainly in English and thereby, to have a distinctive feature in their model they attempted cyberbullying identification in Arabic language. In their work, they utilized various ML algorithms to deal with detecting cyberbullying. The dataset that they gathered had 32 thousand tweets out of which around 1800 tweets were categorised as harassing ones. With the use of algorithms such as Support Vector Machine (SVM), Naïve Bayes they identified cyberbullying and accomplished an accuracy of about 92% and 90% respectively. The outcomes acquired by this framework were not perfect whenever contrasted with past model for English bullying detection. Yet, the point of this paper was to demonstrate that cyberbullying in Arabic language is quite recognizable which shows us that it is very much conceivable to identify harassing in other local or exceptional language too.

PROPOSED SYSTEM

The proposed model comprises of four fundamental segments as demonstrated in Fig. 1

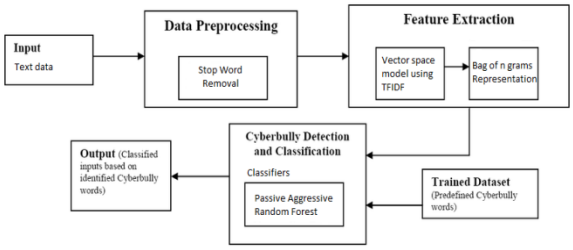

Figure 1: *Architecture Diagram*

The first component comprises of a labeled dataset of social media comments from various platforms in Bengali language. The second component is used for mainly cleaning the data, it basically can be termed as a pre-processing step. Since there might be a possibility that the data gathered may have unstructured content, it is necessary for us to clean or trim the data for obtaining a higher accuracy. The next part of the model is feature extraction. In this segment the preprocessed textual data is changed into an appropriate format in order to prepare the data so that it becomes suitable for machine learning algorithms. The final component of this setup is Classification. Various machine learning techniques or algorithms will be used for examining the training as well as the testing dataset. Performance of all these algorithms will be compared with each other in order to recognize the best suitable one.

IMPLEMENTATION

A. Training and Testing Dataset

In this module, we will focus in on the dataset that we have gathered, eliminating all rows with null entries at first. At

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

that point we will dispose of any unnecessary features that could imperil our algorithm's accuracy. Here we will also divide the dataset into two sections - training and testing. 80% of the dataset will be utilized for training the model and the rest 20% we will use for checking the training models precision. The data gathered is manually labeled as either bully (sexual, threat, troll or religious) or not-bully. Along with it, the dataset also has three other columns specifying the category of the comments passed, to the gender on which the comment is made and total number of reactions for each comment.

B. Data Pre-processing

The data collected had to be preprocessed since it had traces of unstructured contents. It basically meant we needed to clean or trim the data in order to obtain a higher accuracy. There were various steps that were needed to be followed for preprocessing the data such as data cleaning, stop word removal, tokenization. With the help of a stop word filter we deleted any needless words on all the text conversation in line with the Bengali vocabulary. The term stop words mean those words that don't give any helpful data to decide in which category a text should be classified. For facilitating the further processes with the motive of not distinguishing among capital letters and lowercase letters, we transformed the whole data into lower case. Furthermore tokenization had to be practiced on these text contents to facilitate the feature extraction step. Tokenization can be defined as a way of separating or isolating every word that compiles in a document or even a conversation.

	comment	Category	Gender	comment react number	label
0	ওই হালার পুত এখন কি মদ খাওয়ার সময় রাতের বেলা...	Actor	Female	1.0	sexual
1	যার বাস স্টপ করাতে কেমন লেগেছে? ক্যামেরাতে কে ছি...	Singer	Male	2.0	not bully
2	আরে বাবা, এই টা কোন পাগল????	Actor	Female	2.0	not bully
3	ক্যাপ্টেন অফ বাংলাদেশ	Sports	Male	0.0	not bully
4	পটকা মাছ	Politician	Male	0.0	troll

Figure 2: *Dataset Representation*

C. Feature Extraction

The preprocessed data with text conversations will be converted into a vector space model where in these text conversations will be described with a vector of extracted features using Term Frequency Inverse Document Frequency (TFIDF). TFIDF is basically used for measuring or evaluating how relevant a word is to a document or to a collection of documents. Thus, the main aspect of TFIDF is that it performs well on the text and gets the weights of these words regarding the document or the sentence. Along with TFIDF we will also use word level feature extraction methods; this specific strategy is known as Bag of Words or “Bag of n-grams” representation. It implies that documents are defined or represented by occurrences of the words while completely neglecting the position or order of the words in the document. There are various parameter that are mostly used to combine the vectorizer and a machine learning model, one such parameter is max df used to remove terms that appear to frequently in the document.

D. Classification

The final step in the proposed model is classification, where the features extracted are put into an algorithm so as to train and test the classifier and hence to determine whether it can successfully detect cyberbullying or not. We will use various machine learning methods, algorithms such as Support Vector Machine (SVM), Logistic Regression (LR), Random Forest and Passive Aggressive (PR) classifier. The assessment of all these classifiers is completed utilizing few assessment lattices. Among those criteria are Accuracy, precision, recall and f-score.

SVM is a supervised algorithm dependent on discovering a hyper-plane that partitions a given dataset into two different classes. SVM is mainly applicable for text based classification purposes like detecting spam, categorical assignment or even sentimental analysis. SVM is widely used for tasks related to image recognition as well, showing explicitly well outcomes for aspect based recognition and also color based classification.

Logistic regression can be termed as a statistical approach with which we can easily foretell a data input based on previous examinations of the dataset in use. This linear model regression currently has become a vital model in the study of machine learning. This approach basically permits an algorithm which is in use for a machine learning model to classify data based on previous historic data or outcomes. With more of relevant data following, this type of algorithm becomes more precise.

The Passive Aggressive (PA) algorithm is ideal for classifying enormous surges of data (for example Twitter). It's not difficult to carry out and is quick. Passive Aggressive classifier is utilized since there is a huge measure of data and it is computationally infeasible to prepare the whole dataset in light of the sheer size of the information.

Random Forest can be defined as a vigorous or a strong machine learning algorithm which is performed mainly for tasks inclusive of regression and classification. They are described as an ensemble method, implying that these models are made up of many small decision trees, known as estimators, which each can produce its own separate predictions. Random forest in general integrates the outcomes of these estimators to produce a much more accurate result.

RESULT

The below table represents the outcome for all the classifiers that were used in our experiment, the result shows that Passive Aggressive classifier had the highest accuracy when used with N-gram level TFIDF feature extraction. Whereas when using word level feature extraction Support Vector Machine performed the best as compared to others.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

SL NO.	Algorithm Used	Accuracy (in %)	
		Using N-Gram	Using Word Level
1.	Random Forest Classifier	77.1	60.3
2.	Support Vector Machine	77.7	61.1
3.	Logistic Regression	77.5	58.8
4.	Passive Aggressive Classifier	78.1	51.0

Table 1: *Results for the dataset*

CONCLUSION

We made an effort to identify cyberbullying in Bengali language using text classification algorithms. Though we have used various text based classification algorithms such as SVM, Logistic Regression, Passive Aggressive and Random forest but for future purposes, other machine learning models or practices such as CNN and even NLP can be used for the given dataset that we have worked on.

REFERENCES

[1] Abdhullah-Al-Mamun, Shahin Akhter, "Social media bullying detection using machine learning on Bangla text*", 10th International Conference on Electrical and Computer Engineering (ICECE) 2018.*

[2] Md Gulzar Hussain∗ , Tamim Al Mahmud (Member, IEEE), Waheda Akthar, "An Approach to Detect Abusive Bangla Text", *International Conference on Innovation in Engineering and Technology (ICIET) 27-29 December, 2018*

[3] Md Faisal Ahmed, Zalish Mahmud, Zarin Tasnim Biash, Ahmed Ann Noor Ryen, Arman Hossain, Faisal Bin Ashraf R, "Bangla Text Dataset and Exploratory Analysis for Online Harassment Detection" *Department of Computer Science and Engineering, Brac University, 4th February, 2021*

[4] John Hani, Mohamed Nashaat, Mostafa Ahmed, Zeyad Emad, Eslam Amer, Ammar Mohammed, "Social Media Cyberbullying Detection using Machine Learning", *(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 5, 2019*

[5] Rohit Pawar, Rajeev R. Raje, "Multilingual Cyberbullying Detection System", *IEEE International Conference on Electro Information Technology (EIT), 2019*

[6] B. Haidar, M. Chamoun, and F. Yamout, "Cyberbullying detection: A survey on multilingual techniques," *in European Modelling Symposium (EMS), pp. 165–171, Nov 2016.*

[7] Michele Di Capua, Emanuel Di Nardo, and Alfredo Petrosino. "Unsupervised cyber bullying detection in social networks". *In Pattern Recognition (ICPR), 2016 23rd International Conference on, pages 432–437. IEEE, 2016*

[8] Sani Muhamad Isa, Livia Ashianti, et al. "Cyberbullying classification using text mining". *In Informatics and Computational Sciences (ICICoS), 2017 1st International Conference on, pages 241–246. IEEE, 2017.*