Cyber bullying monitoring system for Twitter

Prajakta Ingle [1] Ramya Joshi[1] Neha Kaulgud[1] Aarti Suryawanshi[1] Meghana Lokhande[1]

[1] Pimpri Chinchwad College of Engineering, Pune

DOI: 10.29322/IJSRP.11.04.2021.p11273
http://dx.doi.org/10.29322/IJSRP.11.04.2021.p11273

Abstract: In the recent years Twitter has emerged to be a great source for users to broadcast their daily activities, opinions and feelings via texts and images. Cyberbullying is a harassment that takes prominently happens in social networking sites where cyber bullies target vulnerable victims and it has major psychological and physical effects on the victims. Hence, the Cyberbullying Monitoring System on Twitter is a solution with an aim to identify bullying tweets real time. In our research we have developed a cyberbullying monitoring system.

Index Terms- cyberbullying, light GBM, machine learning.

I. INTRODUCTION

Communication through internet is shifting towards user friendly technologies such as social networking applications ,blogs, online sharing platforms ,etc. The misuse of such technologies leads to cybercrime which includes phishing, malware spreading, spam distribution and cyberbullying [1].Cyberbullying is an ethical issue found on internet and the percentage of the victims is also alarming.

A .CYBERBULLYING

Cyberbullying can be defined as an aggressive or intentionally carried out harassment by group or individual through digital means repeatedly against a sufferer who is unable to defend themself [2].This type of bullying includes threats, abusive or sexual remarks,rumours and hate speech.

B.CYBERBULLYING ON SOCIAL MEDIA SITES

The major contributors to cyberbullying are social networking sites. The dynamic nature of these sites helps in the growth of online aggressive behaviour.The anonymous feature of user profiles increases the complexity to identify the bully. Social media is popular due to its connectivity in the form of networks. But this can be harmful when rumours or bullying posts are spread into the network which cannot be easily controlled. Twitter and Facebook can be taken as examples which are popular among various social media sites.. According to [3] facebook users have more than 150 billion connections which gives the idea about how bullying content can be spread within the network in a fraction of time. To manually identify these bullying messages over this huge network is difficult. There should be an automated system where such kinds of things can be detected automatically thereby taking appropriate action. Researches have shown that cyberbullying has negative effects on victims. The victim mainly consists of women and teenagers [4].Incentive effect on mental and physical health of the victims [2].ease in such kind of activities higher is the risk of depression leading to suicidal cases. Therefore to control cyberbullying there is need of automatic detection or monitoring systems.

C .DETECTION MODELS FOR CYBERBULLYING

Research has shown that machine learning can be used to identify cyberbullying efficiently. We tried to develop a real time cyberbullying monitoring system which will identify bullying and non-bullying tweets. Every time a user selects this option real time tweets will be obtained and classification of these tweets will happen. Along with this analysis of the tweets will be displayed showing the severity and frequency of bullying. Frequency of bullying means how many times a user has bullied others on twitter and severity explains how many bullying words are present in the tweet. These features can be used to take disciplinary actions against such users and can provide analysis for the extent of bullying taking place on twitter. A report will also be displayed showing the results here the date of tweet, the tweet, the twitter handle of the user will be displayed.

II. RELATED WORK

In paper [1], different algorithms are compared for detection of cyberbullying on social media. It consists of comparison between various algorithms like SVM(Support Vector-Machine),NB(naive-Bayes),RF,DT(,KNN(k-nearest-neighbour),LR(logistic-regression),ARM(association-rule-mining),RB(rule based algorithm).Out of all this the SVM algorithm is the best based on factors like accuracy, precision recall .The limitation of this paper is the unexplored deep learning architecture.

In paper [5], there is comparison of Random forest, Naive Bayes, SVM, KNN techniques for detecting and classifying cyberbullying. The most optimal among these all is the Random forest with an average accuracy of 90.8%.The limitations of this paper it suffered oversampling and undersampling which affects the accuracy.

In paper [6], a model was developed to detect cyber bullying .The advantage was that the combination of several algorithms increased the accuracy but this model can be used only for the comment section and it cannot detect homophones.

In paper [7], social media bullying is detected using machine learning. Here supervised machine learning algorithms like SVM with weighted TF IDF was used. Identification of most active predator and victim, improved classification due to Weighted TFIDF.But the limitation was that classification of data must be very accurate, otherwise results won't be accurate

In paper [8], the author generates a model based on pronunciation useful in detection of cyberbullying A rumour detection method based on features of contents which focus on sentiment polarity and opinion of comments, social influence. The limitation is User credibility must be checked for improved results

In paper [9], CNN is used for detection of cyberbullying in twitter. Here training data was labelled using human intelligence service and word embedding was generated for each word using GloVe technique .The resulted set of word embedding was later fed to CNN algorithm for classification. It has a number of strengths like elimination of feature extraction and selection, high accuracy. The limitation of this proposed model is it has language restriction.

On the basis of this survey, we have studied the performance of various machine learning algorithms as well as deep learning algorithms like CNN. Next we implement these as well Light GBM algorithms and results will be shown in the later sections.

III. METHODOLOGY

Selecting a suitable social networking site for detecting cyber bullying is very crucial. Thus Twitter was selected as it generates a lot of data every day and over the years have become a platform for cyberbullying.

Below Figure 1 shows the basic flow of the proposed system

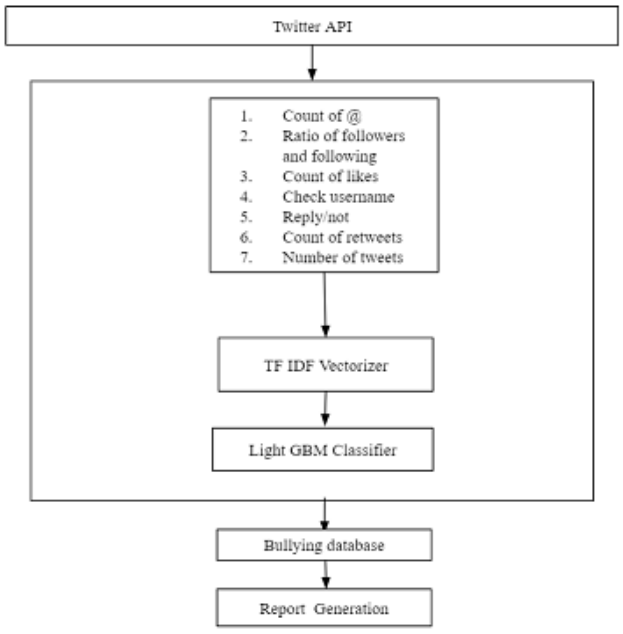

Fig 1: Proposed system

A.DATA COLLECTION:

Twitter API was used to scrape twitter data. As studied in literature survey most of the papers focused on tweet text only. So we have considered some more user specific features like user name, user's no. of following and no. of followers data , total number of tweets user has tweeted and some tweet specific feature like tweet, tweet is a reply or not , number of likes and number of retweets the tweet has. Ratio of followers to following was calculated for implementation. Firstly 4000 tweets were labelled manually as bully or non-bullying tweets. Later the most used slang and abusive words were recognised from manually labelled data. Through this list words term frequency was calculated for labelling of new tweets. Using this method 20,000 bullying and 20,000 non bullying dataset was created with all the mentioned features above.

B.DATA PREPROCESSING:

After data collection all the features except tweet text, username, and label were in string format. For labelling purpose bullying tweets were considered as 1 and 0 for non-bullying. For twitter username the task was to find if the username consisted of any slang words or substring of slang word in it so popular substrings of slang words like ass, fuck, *, $ etc. were matched with the username if any part of username consisted of these substrings then the username was presented as 1 because it contains slang words otherwise 0. The text was cleaned by removing punctuation marks and tags. The count of @tags was stored as a feature while cleaning the text. The whole dataset was shuffled into a 70:30 ratio of train and test datasets. Ratio of followers to following consisted of Nan and infinity values so they were removed.

Tweet text of train and test datasets was converted into numerical data by using TF-IDF vectorizer after tokenization of the sentences. Text was transformed into numeric data using the TFIDF Vectorizer () method. Maximum features were set to 1000. For each text 1000 features were created. These features were concatenated with other remaining features. Label features were separated from train and test datasets and finally the datasets were ready for model implementation..

C. LIGHTGBM

LightGBM provides different gradient boosting algorithms. Gradient boosting machine is used to strengthen weak performing model. Boosting algorithm takes different feature sets and learns from previous predictions and thus selecting best combinations. Gradient boosting decision tree boost the traditional decision tree model. GBDT was used as a classifier using LightGBM library. LightGBM is a library which performs the GBDT algorithm with high efficiency. LightGBM takes less time for execution and works well with large datasets. Model was implemented for 100 iterations for training dataset with learning rate 0.01 and boosting algorithm of gbdt.

D.MONITORING SYSTEM

A web application was developed for getting real time tweets and their classification. On the command of get tweets user is able to scrape real time recent tweets with fixed limit. These tweets are pre-processed and are then given to the saved LightGBM model for classification. The predicted labels along with all user and tweet specific details are stored in the database.

The same database was shown in the form of report through report option. It also consisted of the severity of the bullying tweet. It reflects the percentage of bullying content of a particular tweet.

The analysis of the application provides date wise bullying data; frequency of bullying tweets tweeted by a particular user with user details.

IV. PROJECT SCOPE

In this Cyberbullying Monitoring System for Twitter, it is important to specify the scope in order to accomplish this project efficiently. Since this is a text-based classification system, we have to mention the language and type of text considered for classification. Here, we only focus on English language with proper and formal text. Thus, the informal, abbreviation text, short form and other language will not be considered for classification.

We have developed this system for twitter only and are focusing on the tweets posted by the users in Twitter and consider twitter and user specific features in order to determine the cyberbullying and to display the report and analysis

V. RESULTS

Tab I shows the results of algorithms on the test twitter data set .The table consists of accuracy and confusion matrix of different models. In confusion matrix predicted and actual values of test dataset are represented in form of matrix where 'B' represents bullying tweet and 'NB' represents non bullying tweet. It shows that Light GBM gives the best results.

TABLE 1: COMPARISON OF ALGORITHMS ON TWITTER DATASET

Model name	**Accuracy**	**Confusion matrix**
SVM	98.26	NB B NB 3948 31 B 107 3851
NB	96.63	NB B NB 3798 181 B 86 3872
Logistics Regression:	98.34	NB B NB 3953 26 B 105 3853
Light GBM:	98.67	NB B NB 3978 1 B 104 3854
Decision Tree:	97.85	NB B NB 3903 76 B 94 3864
Random forest	98.66	NB B NB 3973 6 B 100 3858

The table summarizes that Light GBM produces a better result than other algorithms. To provide a better understanding of the model, Fig 2 shows the learning curve of Light-GBM.

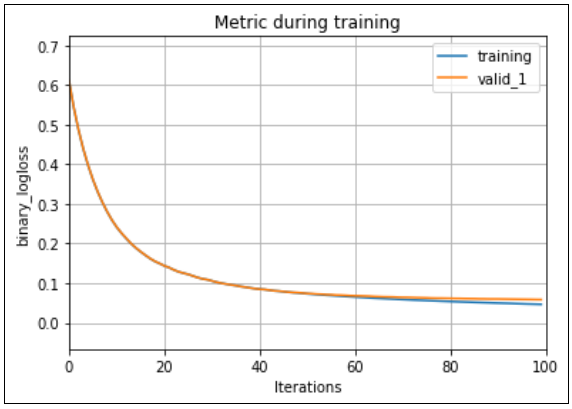

Fig. 2.Learning curve of Light GBM

Fig 3 displays the report of bullying and non-bullying. Here it displays user id, user name, the tweet, result which explains whether it is non-bullying or bullying, date of tweet and severity of bullying words used. This report can give a detailed view of the tweets.

.

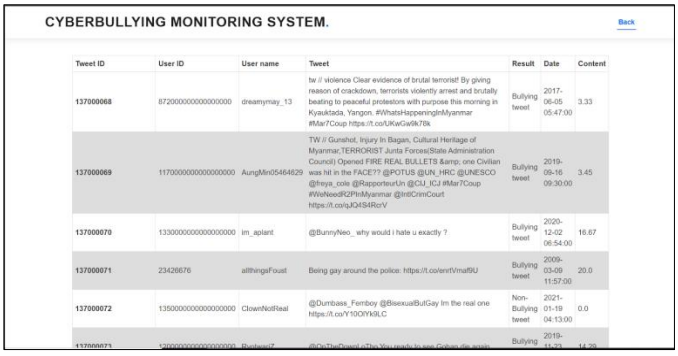

Fig3: Report of tweets

Fig 4 displays the user wise analysis which shows the user-id, user name, followers and following count of users and the number of tweets. Also it shows the number of times that user has posted bullying comments. This can be used to take disciplinary action against such users.

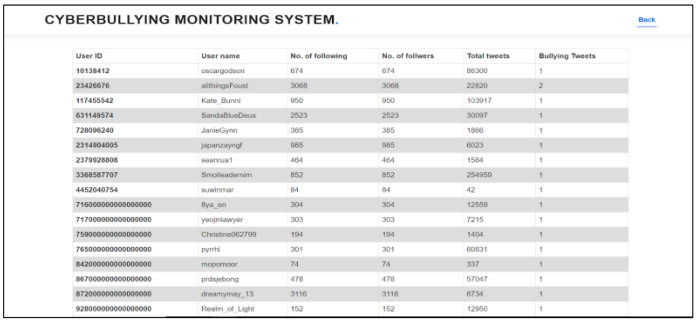

Fig 4: Analysis of tweets

V. CONCLUSION AND FUTURE SCOPE

The study reviewed the existing literature for various machine learning algorithms and identified Light GBM as the most efficient. A model for detecting bullying tweets for real time tweets was developed. We considered various twitter and user specific features along with TF IDF embedding for the classification. A detailed report about tweets and analysis was displayed. In future work, a system to classify these tweets into various categories of bullying can be developed.

REFERENCES

1. M. A. Al-garadi, M. R. Hussain, N. Khan, G. Murtaza, H. F. Nweke, I. Ali, G. Mujtaba, H. Chiroma, H. A. Khattak, and A. Gani, ‘‘Predicting cyberbullying on social media in the big data era using machine learning algorithms:Review-Of-Literature-And-Open-Challenges,’’IEEEAccess,vol.7, pp. 70701–70718, 2019.
2. H. Sampasa-Kanyinga, P. Roumeliotis, and H. Xu, ``Associations between cyberbullying and school bullying victimization and suicidal ideation, plans and attempts among Canadian schoolchildren," PLoS ONE, vol. 9, no. 7, 2014, Art. no. e102145
3. M. Fire, R. Goldschmidt, and Y. Elovici, ``Online social networks: Threats and solutions," IEEE Commun. Surveys Tuts., vol. 16, no. 4, pp. 20192036, 4th Quart., 2014
4. R. Kowalski, G. Giumetti, A. Schroeder and H. Reese, "Cyber Bullying Among College Students: Evidence from Multiple Domains of College Life," Misbehavior Online in Higher Education, vol. 1st,pp. 293-321, 2017
5. .Mohammed Ali Al-garadi, Kasturi Dewi Varathan, Sri Devi Ravana,”Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network” . 0747-5632/© 2016 Elsevier Ltd
6. Society of India CSI (Vol. 2, pp. 637-645). Springer
7. Vikas S Chavan & Shylaja S S (2015).Machine Learning Approach for Detection of Cyber-Aggressive Comments by Peers on Social Media Network
8. Qiao Zhang, Shuiyuan Zhang, Jian Dong, Jinhua Xiong, and Xeuqi Cheng, “Automatic Detection of rumor on Social Network”, pp 113-122, Springer(2015)
9. Monirah A. Al-Ajlan, Mourad Ykhlef, “Optimized Twitter Cyberbullying Detection based on Deep Learning”, 978-15386-4110-1, IEEE-2018
10. E. Raisi, and B. Huang. "Cyberbullying detection with weakly supervised machine learning." In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ACM, 2017, pp. 409-416
11. Chikashi Nobata and Joel Tetreault and Achint Thomas and Yashar Mehdad,” Abusive Language Detection in Online User Content”ACM 978-1-4503-4143-1/16/04.
12. K. Reynolds, "Using Machine Learning to Detect Cyberbullying," the faculty of Ursinus College in fulfillment of the requirements for Distinguished Honors in Computer Science, pp. 1-4, 2012.
13. Vikas S Chavan and Shylaja S S ,”Machine Learning Approach for Detection of Cyber-Aggressive Comments by Peers on Social MediaNetwork”,978-1-4799-8792-4/15/$31.00c 2015 IEEE
14. JVan Royen, K. Poels, W. Daelemans and H. Vandebosch, "Automatic monitoring of cyberbullying on social networking sites: From technological feasibility to desirability," Telematics and Informatics, vol. 32, no. 1, pp. 89-97, 2015.