Machine Learning Based Detecting A Twitter Cyberbullying

V. Gayatri [1], Md. Asim Iqbal [2]

[1] Pursuing M.tech Department of Electronics & Communication Engineering, Kakatiya Univercity College of Engineering &Technology , Warangal, Telangana, India.

[2] Assistant Professor Department of Electronics & Communication Engineering, Kakatiya Univercity College of Engineering &Technology , Warangal, Telangana, India.

[1] vijayapuramgayatri@gmail.com and mdasimiqbal605@gmail.com

Abstract:

Social media is a platform the place many young people are getting bullied. As social networking web sites are increasing, cyberbullying is growing day by using day. To identify word similarities in the tweets made by means of bullies and make use of machine studying and can boost an ML mannequin automatically detect social media bullying actions. However, many social media bullying detection strategies have been implemented, however many of them had been textual based. The intention of this paper is to exhibit the implementation of software program that will notice bullied tweets, posts, etc. A computer getting to know mannequin is proposed to observe and prevent bullying on Twitter. Two classifiers i.e. SVM and Naïve Bayes are used for education and trying out the social media bullying content. Both Naive Bayes and SVM (Support Vector Machine) had been able to notice the genuine positives with 71.25% and 52.70% accuracy respectively. But SVM outperforms Naive Bayes of comparable work on the identical dataset. Also, Twitter API is used to fetch tweets and tweets are handed to the mannequin to observe whether or not the tweets are bullying or not.

Keywords: computing device learning; classifiers: ANN; guide vector computer (SVM); Twitter API

1 Introduction:

Nowadays science has emerge as a very necessary section of our lives and most human beings cannot stay besides it. The Internet provides a platform to share their ideas. Many humans are spending a giant quantity of time on social media. Communicat ing with human beings is no exception, as technology has modified the way humans engage with a broader manner and has given a new dimension to communication. Many people are illegally

the usage of these communities. Many youngsters are getting bullied these days. Bullies use various services like Twitter, Facebook, Email to bully people. Studies show that about 37% of adolescents in India are concerned in cyberbullying and almost 14% of bullying happens regularly. Cyberbullying impacts the sufferer each methods emotionally and psychologically. Social media additionally lets in bullies to harness the anonymity which satisfies their unkind deeds. Things also get extra serious when bullying takes place extra over and over over time. So, stopping it from going on will assist the victim. Cyberbullying and its have an effect on on social media: Cyberbullying is an act of threatening, harassing or bullying someone thru modern-day approaches of communicating with each other and with anybody/everybody in the world with the aid of social media apps/sites. Cyberbullying is now not simply confined to creating a faux identification and publishing/posting some embarrassing photo or video, disagreeable rumors about any person however also giving them threats. The affects of cyberbullying on social media are horrifying, somet imes main to the dying of some unfortunate victims. The conduct of the victims additionally changes due to this, which influences their Emotions, self-confidence and a experience of worry is additionally viewed in such people. Thus, a whole answer is required for this problem. Cyberbullying wishes to stop. The trouble can be tackled by detecting and stopping it by means of the usage of a computing device learning approach, this desires to be accomplished the usage of a specific perspective. The foremost reason of our paper is to advance an ML mannequin so it can become aware of and forestall social media bullying, so no one will have to go through from it. The proposed method is implemented on the social media bullying dataset which used to be accrued from various sources like Kaggle, GitHub, etc. The overall performance of both NB and SVM is in contrast to TFIDF. Twitter API is used to fetch a specific location's tweets to observe whether or not they are Bullying or not. Furthermore, the likelihood of every tweet is calculated to predict the end result and the end result of every tweet is stored into the database with bullies username.

II LITERATURE SURVEY

Detecting social media bullying is carried out by using John Hani et al. [1]. In, their lookup paper, they have used Neural

Networks and classification fashions to observe and stop social media bullying. After doing some research, they sooner or later used NN and SVM for the detection of cyberbullying. For the proposed model they accumulated the dataset from the Kaggle. The proposed mannequin is divided into three foremost steps:

1) Data Preprocessing

2) Feature Extraction

3) Classification

- Preprocessing Steps:

o Tokenization

o Lowering Text

o Stop phrases and encoding cleaning

o Word correction

- Feature Ext raction: For characteristic extraction sentiment analysis and TFIDF algorithms are used.
- Classification: For classification SVM (Support Vector Machine) and NN classifiers are used.

They received higher accuracy the usage of Neural Network classifier i.e. 92.8% and 90.3 p.c the usage of Support Vector Machine whilst the usage of each sentiment evaluation and TFIDF algorithms. Even after comparing their work with preceding work, Neural Network achieved higher accuracy than the Support Vector Machine. For the massive data, Neural Network (NN) performs much better than classification models. Kelly Reynolds et al. [2] has proposed a laptop learning model to observe cyberbullying. In their paper, they've collected facts from Form spring. me internet site the place customers ask and reply the questions. Because of the anonymity of the website, many human beings use it for bullying purposes. Amazon's Mechanical Turk provider is used for labeling information for truth data sets. Data is categorized into two classes. The category label "no" for a tweet besides cyberbullying and "yes" for a t weet with cyberbullying. Machine Learning algorithms have been used for predicting attributes and statistics sets. Two different training units have been extracted one for counting information and one for normalizing the informat ion. J48, JRIP, IBK, AND SMO ALGORITHMS have been used for education sets. J48 is used for developing a choice tree. Interestingly overall the received accuracy used to be 81.7%. Amanpreet Singh et al. [3] has reviewed many previous research papers associated to laptop getting to know models,

preprocessing techniques, comparison of desktop learning models, etc. This paper consists of find out about lookup based totally on various preceding lookup papers. They've mentioned used methodology,datasets,conclusions/findings, content-based features, demerits, method and used models, preprocessing steps used for the model. For, getting to know purposes, they've explored Scopus and the IEEE Xplore digital library, ACM Digital Library. Using citations, fifty one educational papers were discovered. Based on concluding arguments, abstracts, and titles, 18 papers had been located no longer to observe to the survey so 18 papers have been discarded. In this paper for the survey, they've reviewed 27 papers from 33 papers after filtration. In, every of the 27 lookup papers binary classification is used for cyberbullying detection. And most of them have used the Support Vector Machine (SVM) algorithm for detection. Abdhullah-Al-Mamunet al. [4] has developed a machine learning mannequin to realize social media bullying for Bangla text. In this paper, they are detecting cyberbullying for Bangla text. For this, they've proposed a number laptop learning algorithms for cyberbullying detection on Bangla text. To develop a mannequin for the Bangla textual content dataset has been collected from a number of social media structures (such as Facebook Graph API and Twitter REST API) and for t raining purposes, labeled them both bullied or no longer bullied. They have used supervised Machine Learning algorithms i.e. SVM, KNN, and NB (Naive Bytes) classifier models.

Accuracy of each and every Model:
SVM-97.27% | KNN-96.73 |
NB- 97.23

SVM outperformed different classifiers for each English and Bangla text. Moving to the subsequent paper, the Support Vector Machine (SVM) algorithm is used by means of Potha et al. [5] and they additionally achieved 49.8% accuracy. Moreover, SVM and Lexical Syntactic approach for function extract ion had been used through Chen et al. [6] to detect abusive language and they obtained 77.9% precision value. Now, transferring to the subsequent paper proposed via Chavan et al. [7] has used SVM and Logistic Regression for the classification of the data, with SVM they obtained 77.65% accuracy and 73.76% using logistic regression. Furthermore, in the subsequent strategy by Romsaiyud et al. [8], they

have improvised the Naïve Bayes [NB] c lassifier fo r e xt racting the phrases and inspected thoroughly the loaded sample clustering. While implementing this method, an accuracy of 95.79% is finished on certain datasets from Slashdot, Kongregate, MySpace, etc. But there is a issue to this method due to the fact the clustering processes do now not coordinate with every different whilst working. As in all the preceding research, fashions have been developed are no longer carried out on any real-time data. Very few works have been completed on actual t ime-data, so laptop getting to know models are applied on Twitter's real-t ime tweets the usage of Twitter API

III Proposed Method

In this paper, a answer is proposed to observe twitter cyberbullying. The most important distinction with preceding lookup is that we now not solely developed a laptop gaining knowledge of mannequin to detect cyberbullying content material however additionally applied it on particularlocations real-time tweets the use of Twitter API. The whole strategy to observe and forestall Twitter cyberbullying is divided into two primary stages: creating the model and experimental setup.

1. Experimental Setup:

Stepwise Procedure of SVM and Naïve Bayes utilized in detecting the cyberbullying Steps:

1. For a specific location, a restrained variety of tweets will be fetched thru Twitter's tweet API [10]
2. The Data Preprocessing, Data Ext raction will be performed on the fetched Tweets
3. Preprocessed tweets will be exceeded to SVM and Naïve Bayes mannequin (see Developing the Model section) to calculate the chances of fetched tweets to check whether a fetched tweet is bullying or not.

4. If the chance of fetched tweet lies in the vary of zero to 0.5, then the tweet will no longer be regarded as a bullied tweet. If the chance of the fetched tweet is above 0.5, it will be introduced to the database and then similarly 10 tweets from that users' timeline will be fetched, due to the fact it cannot directly say the man or woman is bullying any person or no longer because it is may viable he is having a dialog with his friend for this reason to make certain whether or not he was once bullying someone or now not we will fetch final 10 tweets from his timeline

and preprocessing will be carried out over the tweets.

5. Again, the listing of user's time line t weets will be handed to the SVM and Naive Bayes mannequin to predict the outcomes of the tweets.

6. And again, the common chance of that user's tweets will be calculated and if it lies above 0.5 then it will be considered as a bullied tweet and it will be recorded in our database. If the common likelihood is much less than 0.5 then the file will be eliminated from the database.

Fig. I exhibit the flowchart of the proposed solution. The first step in the answer is to accumulate the tweets from Twitter using Twitter API. In the subsequent two steps are records preprocessing and feature extraction is carried out over the tweets. And after performing preprocessing and function extract ion tweets are passed to the SVM mannequin for classificat ion to predict whether the tweet is Bullying or Non-Bullying.

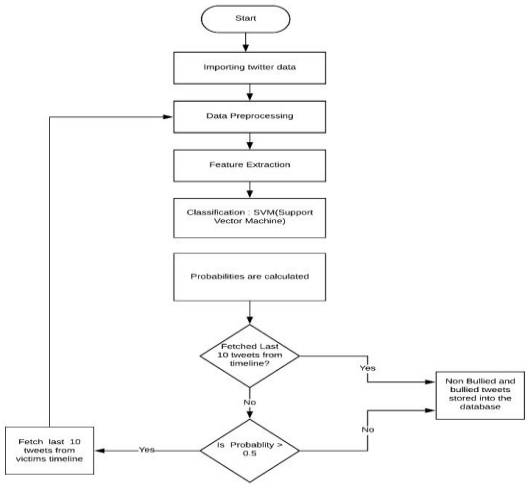

Fig. I: Flowchart of the ent ire experimental setup

2. Developing the Model:

The entire model is divided into 3 major steps: Preprocessing, the algorithm, and feature extraction.

A. Preprocessing:

The Natural Language Toolkit (NLTK) is used for the preprocessing of data. NLTK is used for tokenization of text patterns, to remove stop words from the text, etc.

 Tokenization: In tokenizat ion, the input text is split as the separated words and words are appended to the list. Firstly, PunktSentenceTokenizer is used to tokenized text into the sentences [11]. Then 4 different tokenizers are used to tokenize the sentences into the words:

o WhitespaceTokenizer

o WordPunctTokenizer

o TreebankWordTokenizer

o PunctWordTokenizer

 Lowering Text : It lowers all the letters of the words from the tokenization list. Example: Before lowering “Hey There” after lowering “hey there”.

 Removing Stop words: This is the most important part of the preprocessing. Stop words are useless words in the data. Stop words can be get rid of very easily using NLTK. In this stage stop words like \t, https, \u, are removed from the text.

 Wordnet lemmatizer: Wordnet lemmat izer finds the synonyms of a word, meaning and many more and links them to the one word.

B. Feature Extraction:

In this step, the proposed model has transformed the data in a suitable form which is passed to the machine learning algorithms. The TFDIF vectorizer [1] is used to extract the features of the given data. Features of the data are extracted and put them in a list of features. Also, the polarity (i.e. the text is Bullying or Non-Bullying) of each text is extracted and stored in the list of features.

C. Algorithm Selection:

To detect social media bullying automatically, supervised Binary classification machine learning algorithms like SVM with linear kernel and Naive Bayes is used. The reason behind this is both SVM and Naive Bayes calculate the probabilities for each class (i.e. probabilities of Bullying and Non-Bullying tweets). Both SVM and NB algorithms are used for the classification of the two-cluster. Both the machine learning models were evaluated on the same dataset. But SVM outperformed Naive Bayes of similar work on the same dataset. Classification report [9] is also evaluated. The accuracy, recall, f-score, and precision are also calculated.

Precision = TP / (TP+FP)

Recall =TP/(TP+FN)

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

Where TP = True positive numbers

TN = True bad numbers

FN = False terrible numbers

FP = False high quality numbers

I) SVM (Support Vector Machine)

Support Vector Machine is a supervised classification machine learning algorithm. SVM can be used for each regression and classification. SVM additionally calculates the probabilit ies for each category [12]. SVM with Linear Kernel is used as our records is linearly separable.

HYPERPLANE:

The most important purpose of the SVM

is to locate the hyperplane which divides the dataset into two categories. Many hyperplanes separate two classes of the records points. The foremost goal of the SVM is to discover the hyperplane with a most margin. For 2 attributes hyperplane is simply a line. As the quantity of features increases, it is very difficu lt to think about the hyperplanes' dimension. In our model, as there are solely two lessons , i.e. Bullying and Non-Bullying hyperplane used to be simply a line.

SUPPORT VECTORS:

Data factors that are nearer to the hyperplane are known as Support Vectors. To maximize the marginal distance between classifiers guide vectors are used and if delete this support vector it will exchange the hyperplanes' position.

II) Naive Bayes

Naive Bayes is a supervised probabilistic computer learning algorithm that can be used for classification [13]. Bayes Theorem Formula: Naive Bayes fashions are used suggestion systems, sentiment analysis, and unsolicited mail filtering. Naive Bayes algorithms are very handy to implement.Types of Classifiers:

- Gaussian Naive Bayes
- Bernoulli Naive Bayes
- Multinomial Naive Bayes

Since our records is no longer discrete Gaussian Naïve Bayes approach is used.

IV RESULT

In this section, the SVM and Naive Bayes on the dataset collected from the various sources like Kaggle, Github, etc are compared. After performing preprocessing and feature extraction on the dataset, for training and testing, and divided the dataset into ratios 0.45 and 0.55 respectively. Both SVM and Naive Bayes are evaluated to calculate the accuracy, recall, f-score, and precision. Interestingly SVM outperformed Naive Bayes in every aspect. Table I shows the accuracies of both the Naive Bayes and SVM. The Support Vector Machine achieved the highest accuracy i.e. 71.25%, while Naive Bayes achieved 52.70% accuracy. Fig. II shows both classifiers accuracy results. Table III shows that the SVM algorithm achieved the highest precision value i.e. 71%, while NB achieved 52% precision. Also, SVM has achieved higher recall and f-score values than Naive Bayes. Fig. III shows the results of the experimental setup, where tweets are fetched from

Twitter using Twitter API with the username of the person and Fig. IV shows the result of the fetched tweets that is whether the tweets are bullying or not with their probability, where the 7th tweet is detected as a “Bullying” tweet and rest are “Non-Bullying”. Fig. V shows the final result with the average probability of bul lied tweet which is reduced to 0.15 from 0.54, so the 7th is labeled as a “Non-Bullying” tweet.

Classifiers	Accuracy (in %)
Naive Bayes	52.70
Support Vector Machine	71.25

TABLE I: The Accuracy of Support Vector Machine and Naive Bayes

Classifiers	Precision	Recall	F-Score
Naive Bayes	52%	52%	53%
Support Vector Machine	71%	71%	70%

TABLE II: Classificat ion Report of the Naive Bayes Algorithm

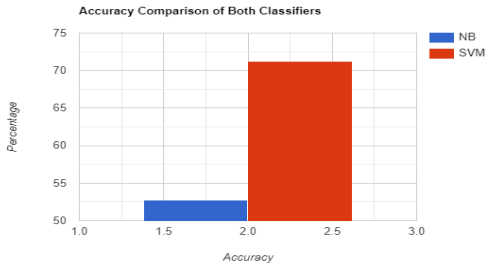

Fig. II: Accuracy Comparison of Both Classifiers

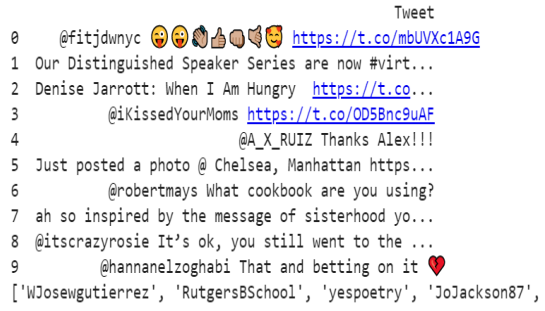

Fig. III: Fetched Tweets Using Twit ter API

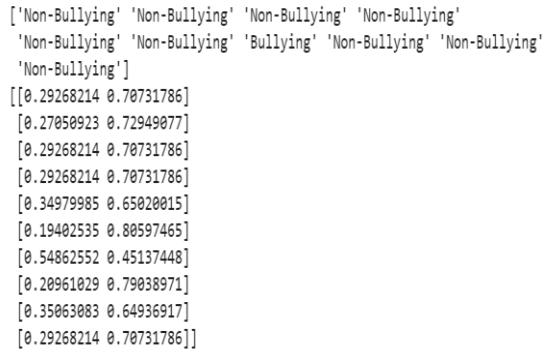

Fig. IV: Results of Fetched Tweets

V FUTURE SCOPE

As this mission predict bullying with the aid of pasting the tweet in the search bar and the detecting bullied or now not and as for the future scope it can additionally be developed via including points which will become aware of in the actual time. By commenting tweet in the Twitter account, it will observe the tweet as the harassment one and will no longer let the attacker to tweet it. By including new points in the Twitter application, it can be in addition modified to use as a herbal social media app. As in this mission it is

utilized in the Twitter, so it can be in addition be utilized in a range of social media app.

VI CONCLUSION

An method is proposed for detecting and stopping Twitter cyberbullying the use of Supervised Binary classification Machine Learning algorithms. Our mannequin is evaluated on each Support Vector Machine and Naive Bayes, additionally for characteristic extract ion, used the TFIDF vectorizer. As the outcomes exhibit us that the accuracy for detecting cyberbullying content material has additionally been great for Support Vector Machine of round 71.25% which is better than Naive Bayes. Our model will assist human beings from the attacks of social media bullies.

VI References

[1] John Hani Mounir, Mohamed Nashaat, Mostafa Ahmed, Zeyad Emad, Eslam Amer, Ammar Mohammed, “ Social Media Cyberbullying Det ect ion using Machine Learning”, (IJACSA) Internat ional Journal of Advanced Computer Science and Applications Vol. 10, pages 703-707, 2019.

[2] Kelly Reynolds, April Kontostathis, Lynne Edwards, "Using Machine Learning t o Detect Cyberbullying”, 2011 10t h Int ernat ional Conference on Machine Learning and Applications volume 2, pages 241–244. IEEE, 2011

[3] Amanpreet Singh, Maninder Kaur, "Content -based Cybercrime Detect ion: A Concise Review”, Internat ional Journal of Innovat ive Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-8, pages 1193-1207, 2019

[4] Abdhullah-Al-Mamun, Shahin Akhter, "Social media bullying detect ion using machine learning on Bangla t ext ”, 10th Internat ional Conference on Elect rical and Computer Engineering, pages 385-388, IEEE Xplore, 2018

[5]Nekt ariaPot ha and ManolisMaragoudakis. “ Cyberbullying det ectionusing time series modeling”, In 2014IEEE Int ernat ional Conference on, pages 373– 382. IEEE, 2014.

[6] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. “Det ect ing offensive language in social media to prot ect adolescent online safety”. In P rivacy, Security, Risk and Trust (PASSAT), 2012 Internat ional Conference on and 2012

Internat ional Conference on Social Computing (SocialCom), pages 71– 80. IEEE, 2012

[7] Vikas S Chavan, SS Shylaja. “Machine learning approach for det ect ion of cyber-aggressive comments by peers on social media network". In Advances in comput ing, communicat ions, and informat ics (ICACCI), 2015 Internat ional Conference on, pages 2354– 2358. IEEE,2015

[8] Walisa Romsaiyud, Kodchakorn na Nakornphanom, Pimpaka Prasertsilp, P iyaporn Nurarak, and P irom Konglerd, “Automated cyberbullying detect ion using clust ering appearance patterns”, In Knowledge and Smart Technology (KST), 2017 9th International Conference on, pages 242– 247. IEEE, 2017.

[9] ht tps://muthu.co/understanding-the-classification-report-in-sklearn/

[10] https://developer.twit ter.com/en/apps

[11]https://textprocessing.com/demo/tokenize/

[12]https://towardsdatascience.com/support -vector-machine-int roduct ion tomachine-learning-algorithms-934a444fca47

[13] https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c