Hate Speech Detection System Using Deep Learning And NLP

Bhavana K[1], Nisarga R[2], Sahaana V[3], Vaishali P[4]
[1]*Bhavana K, Dept. of AD, East West Institute of Technology, Bengaluru,*
[2]*Nisarga R, Dept. of AD, East West Institute of Technology, Bengaluru,*
[3]*Sahaana V, Dept. of AD, East West Institute of Technology, Bengaluru,*
[4]*Vaishali P, Dept. of AD, East West Institute of Technology, Bengaluru,*

Abstract— Social media is a platform where many young people are getting bullied. As social networking sites are increasing, cyberbullying is increasing day by day. To identify word similarities in the tweets made by bullies and make use of machine learning and can develop an ML model automatically detect social media bullying actions. However, many social media bullying detection techniques have been implemented, but many of them were textual based. The goal of this paper is to show the implementation of software that will detect bullied tweets, posts, etc. A machine learning model is proposed to detect and prevent bullying on Twitter. Two classifiers i.e. SVM and Naïve Bayes are used for training and testing the social media bullying content. Both Naive Bayes and SVM (Support Vector Machine) were able to detect the true positives with 71.25% and 52.70% accuracy respectively. But SVM outperforms Naive Bayes of similar work on the same dataset..

Keywords- cyberbullying; machine learning; classifiers; Naive Bayes; support vector machine (SVM)

I. INTRODUCTION

Nowadays technology has become a very important part of our lives and most people can't live without it. The Internet provides a platform to share their ideas. Many people are spending a large amount of time on social media. Communicating with people is no exception, as technology has changed the way people interact with a broader manner and has given a new dimension to communication. Many people are illegally using 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 children in India are involved in cyberbullying and nearly 14% of bullying occurs regularly. Cyberbullying affects the victim both ways emotionally and psychologically. Social media also allows bullies to harness the anonymity which satisfies their unkind deeds. Things also get more serious when bullying occurs more repeatedly over time. So, preventing it from happening will help the victim.

Cyberbullying and its impact on social media:
Cyberbullying is an act of threatening, harassing or bullying someone through modern ways of communicating with each other and with anybody/everybody in the world via social media apps/sites. Cyberbullying is not just limited to creating a fake identity and publishing/posting some embarrassing photo or video, unpleasant rumors about someone but also giving them threats. The impacts of cyberbullying on social media are horrifying, sometimes leading to the death of some unfortunate victims. The behavior of the victims also changes due to this, which affects their Emotions, self-confidence and a sense of fear is also seen in such people.

Thus, a complete solution is required for this problem. Cyberbullying needs to stop. The problem can be tackled by detecting and preventing it by using a machine learning approach, this needs to be done using a different perspective. The main purpose of our paper is to develop an ML model so it can detect and prevent social media bullying, so nobody will have to suffer from it. The proposed technique is implemented on the social media bullying dataset which was collected from various sources like Kaggle, GitHub, etc. The performance of both NB and SVM is compared to TFIDF. Twitter API is used to fetch a particular location's tweets to detect whether they are Bullying or not. Furthermore, the probability of each tweet is calculated to predict the result and the result of each tweet is stored into the database with bullies username.

II. LITERATURE SURVEY

Detecting social media bullying is done by John Hani et al. [1]. In, their research paper, they have used Neural Networks and classification models to detect and prevent social media bullying. After doing some research, they finally used NN and SVM for the detection of cyberbullying. For the proposed model they collected the dataset from the Kaggle. The proposed model is divided into 3 major steps: 1) Data Preprocessing 2) Feature Extraction 3) Classification

- Preprocessing Steps: ○ Tokenization ○ Lowering Text ○ Stop words and encoding cleaning ○ Word correction
- Feature Extraction: For **feature extraction** sentiment analysis and TFIDF algorithms are used.
- Classification: For **classification** SVM (Support Vector Machine) and NN classifiers are used. They got better accuracy using Neural Network classifier i.e. 92.8% and 90.3 % using Support Vector Machine while using both sentiment analysis and TFIDF algorithms. Even after comparing their work with previous work, Neural Network achieved better accuracy than the Support Vector Machine. For the large data, Neural Network (NN) performs much better than classification models.

Kelly Reynolds et al. [2] has proposed a machine learning model to detect cyberbullying. In their paper, they've collected data from Formspring.me website where users ask and answer the questions. Because of the anonymity of the website, many people use it for bullying purposes. Amazon's Mechanical Turk service is used for labeling data for truth data sets. Data is classified into two classes. The class label "no" for a tweet without cyberbullying and "yes" for a tweet with cyberbullying. Machine Learning algorithms have been used for predicting attributes and data sets. Two different training sets have been extracted one for counting information and one for normalizing the information. J48, JRIP, IBK,
AND SMO ALGORITHMS have been used for training sets. J48 is used for creating a decision tree. Interestingly overall the obtained accuracy was 81.7%.

Amanpreet Singh et al. [3] has reviewed many previous research papers related to machine learning models, preprocessing techniques, evaluation of machine learning models, etc. This paper includes study research based on various previous research papers. They've discussed used methodology, datasets, conclusions/findings, content-based features, demerits, technique and used models, preprocessing steps used for the model. For, researching purposes, they've explored Scopus and the IEEE Xplore virtual library, ACM Digital Library. Using citations, 51 academic papers were discovered. Based on concluding arguments, abstracts, and titles, 18 papers were found not to apply to the survey so 18 papers were discarded. In this paper for the survey, they've reviewed 27 papers from 33 papers after filtration. In, each of the 27 research 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 model to detect social media bullying for Bangla text. In this paper, they are detecting cyberbullying for Bangla text. For this, they've proposed various machine learning algorithms for cyberbullying detection on Bangla text. To develop a model for the Bangla text dataset has been collected from various social media platforms (such as Facebook Graph API and Twitter REST API) and for training purposes, labeled them either bullied or not bullied.
They have used supervised Machine Learning algorithms i.e.
SVM, KNN, and NB (Naive Bytes) classifier models. Accuracy of every Model:
SVM-97.27% | KNN-96.73 | NB- 97.23
SVM outperformed other classifiers for both English and Bangla text.

Moving to the next paper, the Support Vector Machine (SVM) algorithm is used by Potha et al. [5] and they also achieved 49.8% accuracy. Moreover, SVM and Lexical Syntactic approach for feature extraction were used by Chen et al. [6] to detect abusive language and they got 77.9% precision value. Now, moving to the next paper proposed by Chavan et al. [7] has used SVM and Logistic Regression for the classification of the data, with SVM they got 77.65% accuracy and 73.76% using logistic regression. Furthermore, in the next approach by Romsaiyud et al. [8], they have improvised the Naïve Bayes [NB] classifier for extracting the words and inspected thoroughly the loaded pattern clustering. While implementing this method, an accuracy of 95.79% is achieved on certain datasets from Slashdot, Kongregate, MySpace, etc. But there is a limitation to this approach because the clustering processes do not coordinate with each other while working.

As in all the previous research, models have been developed are not implemented on any real-time data. Very few works have been done on real time-data, so machine learning models are implemented on Twitter's real-time tweets using Twitter API

III. PROPOSED SOLUTION

In this paper, a solution is proposed to detect twitter cyberbullying. The main difference with previous research is that we not only developed a machine learning model to detect cyberbullying content but also implemented it on particular locations real-time tweets using Twitter API.
The entire approach to detect and prevent Twitter cyberbullying is divided into 2 major stages: developing the model and experimental setup.

1. Experimental Setup:

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

Steps:

1. For a particular location, a limited number of tweets will be fetched through Twitter's tweet API [10]

2. The Data Preprocessing, Data Extraction will be performed on the fetched Tweets

3. Preprocessed tweets will be passed to SVM and Naïve Bayes model (see Developing the Model section) to calculate the probabilities of fetched tweets to check whether a fetched tweet is bullying or not.

4. If the probability of fetched tweet lies in the range of 0 to 0.5, then the tweet will not be considered as a bullied tweet. If the probability of the fetched tweet is above 0.5, it will

5. be added to the database and then further 10 tweets from that users' timeline will be fetched, because it cannot

6. directly say the person is bullying someone or not because it is might possible he's having a conversation with his friend hence to make sure whether he was bullying someone or not we will fetch last 10 tweets from his timeline and preprocessing will be performed over the tweets.

7. Again, the list of user's timeline tweets will be passed to the SVM and Naive Bayes model to predict the results of the tweets.

8. And again, the average probability 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 average probability is less than 0.5 then the record will be removed from the database.

Fig. I show the flowchart of the proposed solution. The first step in the solution is to collect the tweets from Twitter using Twitter API. In the next two steps are data preprocessing and feature extraction is performed over the tweets. And after performing preprocessing and feature extraction tweets are passed to the SVM model for classification to predict whether the tweet is Bullying or Non-Bullying.

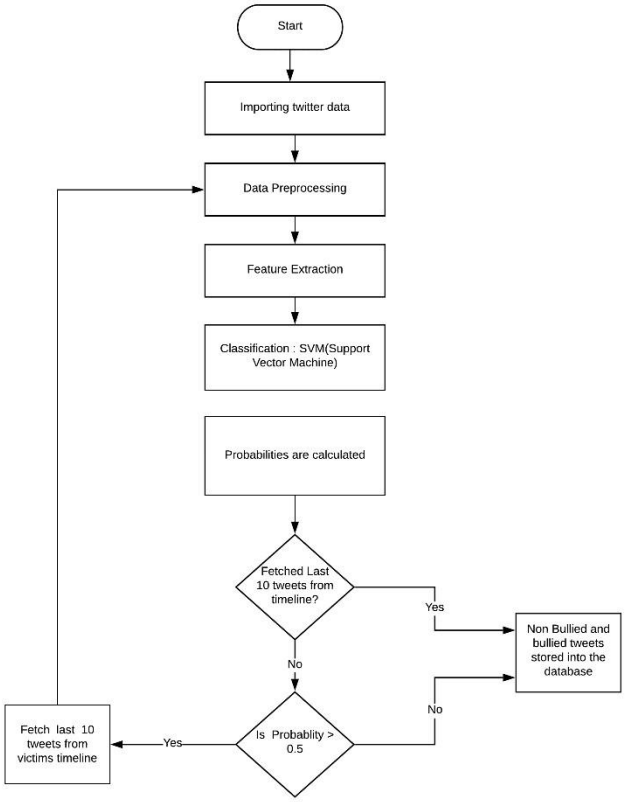

Fig. I: Flowchart of the entire experimental setup

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 tokenization, 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:
  - WhitespaceTokenizer
  - WordPunctTokenizer
  - TreebankWordTokenizer
  - 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 lemmatizer 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 negative numbers

FN = False negative numbers

FP = False positive numbers

I) SVM (Support Vector Machine)

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

HYPERPLANE:

The main aim of the SVM is to find the hyperplane which divides the dataset into two categories. Many hyperplanes separate two categories of the data points. The main aim of the SVM is to find the hyperplane with a maximum margin. For 2

attributes hyperplane is just a line. As the number of features increases, it is very difficult to imagine the hyperplanes' dimension. In our model, as there are only 2 classes, i.e. Bullying and Non-Bullying hyperplane was just a line.

SUPPORT VECTORS:
Data points that are closer to the hyperplane are called Support Vectors. To maximize the marginal distance between classifiers support vectors are used and if delete this support vector it will change the hyperplanes' position.

II) Naive Bayes

Naive Bayes is a supervised probabilistic machine learning algorithm that can be used for classification [13]. Bayes Theorem Formula:

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

A, B = events
$P(A|B)$ = probability of A given B is true
$P(B|A)$ = probability of B given A is true
$P(A), P(B)$ = the independent probabilities of A and B

Naive Bayes models are used recommendation systems, sentiment analysis, and spam filtering. Naive Bayes algorithms are very easy to implement. Types of Classifiers:

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

Since our data is not 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 bullied 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: Classification Report of the Naive Bayes Algorithm

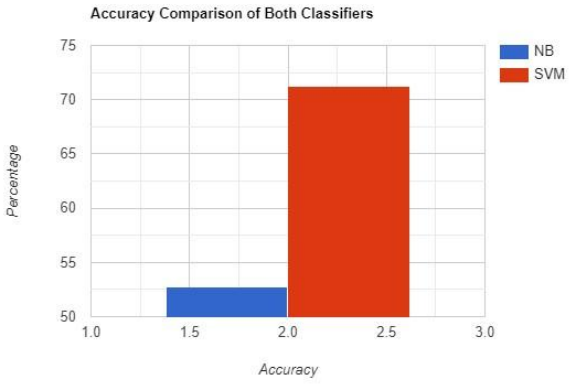

Fig. II: Accuracy Comparison of Both Classifiers

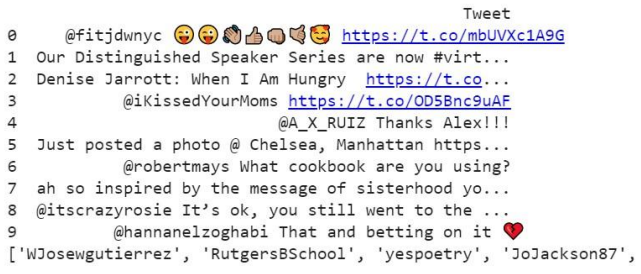

Fig. III: Fetched Tweets Using Twitter API

```
['Non-Bullying' 'Non-Bullying' 'Non-Bullying' 'Non-Bullying'
 'Non-Bullying' 'Non-Bullying' 'Bullying' 'Non-Bullying' 'Non-Bullying'
 'Non-Bullying']
[[0.29268214 0.70731786]
 [0.27050923 0.72949077]
 [0.29268214 0.70731786]
 [0.29268214 0.70731786]
 [0.34979985 0.65020015]
 [0.19402535 0.80597465]
 [0.54862552 0.45137448]
 [0.20961029 0.79038971]
 [0.35063083 0.64936917]
 [0.29268214 0.70731786]]
```

Fig. IV: Results of Fetched Tweets

```
['Nick_Brown10']
iiii @hannanelzoghabi That and betting on it 💔
1
[ 0.15869234624889328]
['Nick_Brown10']
```

Fig. V: Final Result

V. CONCLUSION

An approach is proposed for detecting and preventing Twitter cyberbullying using Supervised Binary classification Machine

Learning algorithms. Our model is evaluated on both Support Vector Machine and Naive Bayes, also for feature extraction, used the TFIDF vectorizer. As the results show us that the accuracy for detecting cyberbullying content has also been great for Support Vector Machine of around 71.25% which is better than Naive Bayes. Our model will help people from the attacks of social media bullies.

REFERENCES

[1] John Hani Mounir, 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, pages 703-707, 2019.

[2] Kelly Reynolds, April Kontostathis, Lynne Edwards, "Using Machine Learning to Detect Cyberbullying”, 2011 10th International Conference on Machine Learning and Applications volume 2, pages 241–244. IEEE, 2011

[3] Amanpreet Singh, Maninder Kaur, "Content-based Cybercrime Detection: A Concise Review”, International Journal of Innovative 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 detection using machine learning on Bangla text”, 10th International Conference on Electrical and Computer Engineering, pages 385-388, IEEE Xplore, 2018 [5]NektariaPotha and ManolisMaragoudakis. “Cyberbullying detectionusing time series modeling”, In 2014IEEE International Conference on, pages 373– 382. IEEE, 2014.

[6] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. “Detecting offensive language in social media to protect adolescent online safety”. In Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom), pages 71–80. IEEE, 2012

[7] Vikas S Chavan, SS Shylaja. “Machine learning approach for detection of cyber-aggressive comments by peers on social media network". In Advances in computing, communications, and informatics (ICACCI), 2015 International Conference on, pages 2354–2358. IEEE,2015

[8] Walisa Romsaiyud, Kodchakorn na Nakornphanom, Pimpaka Prasert- silp, Piyaporn Nurarak, and Pirom Konglerd, “Automated cyberbullying detection using clustering appearance patterns”, In Knowledge and Smart Technology (KST), 2017 9th International Conference on, pages 242–247. IEEE, 2017.

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

[10] https://developer.twitter.com/en/apps

[11] https://text-processing.com/demo/tokenize/

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

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