ZKG INTERNATIONAL ISSN: 2366-1313

MACHINE LEARNING BASED TWITTER CYBER BULLYING DETECTION

[1] KORUKONDA VENKATA RATHNAM, [2] PANCHADARLA APOORVA, [3] MEDABALIMI CHAKRAVARTHI, [4] ISMAIL SYED, [5] PALEM SIVA MAHESWARA REDDY

[1]Assistant Professor, Dept. Of CSE, ABR College of Engineering and Technology, Kanigiri

[2, 3, 4, 5,] BTech Student, Dept. Of CSE, ABR College of Engineering and Technology, Kanigiri

Abstract: *In today's digital society, cyberbullying has become grave issue affecting an increasingly high number of Internet users, mostly at their sensitive teen and young age on social media platforms such as Instagram. Most of the bullying involves intimidation or mean comments that focus on things like a person's gender, religion, sexual orientation, race, or physical differences count as discrimination, which is against the law in many states. Cyberbullying is a Psychological Abuse which leads to mental abuse. Thus, to reduce this we have chosen our project to detect cyberbullying comments in Twitter. 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. Also, Twitter API is used to fetch tweets and tweets are passed to the model to detect whether the tweets are bullying or not.*

Keywords: *Cyberbullying, social media, BERT, NLP, Machine learning, Twitter*

I. INTRODUCTION

Millions of young people spend their time on social networking, and the sharing of information is online. Social networks have the ability to communicate and to share information with anyone, at any time, and in the number of people at the same time. There are over 3 billion social media users around the world. According to the National Crime Security Council (NCPC), cyberbullying is available online where mobile phones, video game apps, or any other way to send or send text, photos, or videos deliberately injure or embarrass another person. Cyberbullying can happen at any time all day, week and you can reach anyone anywhere via the internet [1]. Text, photos, or videos of cyberbullying

may be posted in an undisclosed manner. It can be difficult, and sometimes impossible, to track down the source of this post. It was also impossible to get rid of these messages later. Several social media platforms such as Twitter, Instagram, Facebook, YouTube, Snapchat, Skype, and Wikipedia are the most common bullying sites on the internet. Some of the social networking sites, such as Facebook, and the provision of guidance on the prevention of bullying. It has a special section that explains how to report cyber-bullying and to prevent any blocking of the user. On Instagram, when someone shares photos and videos made by the user to be uncomfortable, so the user can monitor or block them. Users can also report a violation of our community and make Recommendations to the app. While these platforms provide an opportunity for people to interact and communicate in ways that were previously unimaginable, they have also given rise to negative behaviours like cyberbullying. Cyberbullying is the act of intimidating, threatening, or coercing others through the internet using digital or electronic means such as social media, email, text messaging, blog postings. Cyberbullying, also known as internet harassment, frequently makes use of insulting, hostile, or threatening language. Cyberbullies frequently hide their true identities behind fake digital profiles [2].

Cyberbullying is a major and widespread problem in today's digital culture that affects a growing number of Internet users, particularly impressionable teenagers and young people. In a way, unlike its digital equivalent, which can happen anytime, anywhere with only a few keystrokes on a keyboard, physical bullying is relatively restricted to specific locations or periods of the day.

Cyberbullying is a form of psychological abuse that has a big influence on society. Events of cyberbullying have been rising, especially among young individuals who spend the majority of their time switching between various social media sites. Because of their popularity and the anonymity that the Internet offers to abusers, social media networks like Twitter and Instagram are particularly vulnerable. Cyberbullying may even result in severe mental disorders and detrimental impacts on mental health. The majority of suicides are caused by the worry, depression, stress, and social and emotional challenges brought on by instances of cyberbullying [3].

These issues lead to the creation of techniques and tools for the early identification and prevention of such

abusive behaviour, particularly when it develops on social media platforms. Developing efficient and effective strategies for detecting such online occurrences involves many complexities. This highlights the need for a method to spot cyberbullying in messages posted on social media (e.g., posts, tweets, and comments). The key tasks in addressing cyberbullying risks are the detection of cyberbullying events from tweets and the implementation of preventive measures. This is because cyberbullying is increasingly an issue on Instagram. Therefore, there is a larger need to conduct more study on social network-based CB in order to gain more knowledge and contribute to the creation of tools and strategies that will successfully tackle the problem.

The main methods for detecting cyberbullying on the Instagram platform are comment categorization and, to a lesser extent, topic modelling techniques. Text categorization using supervised machine learning (ML) models is frequently used to separate bullying-related and non-bullying comments. Bullying and non-bullying tweet classification has also been accomplished using deep learning (DL) based classifiers. Only a predetermined set of events may be adequate for supervised classifiers; however, they are unable to handle dynamically changing comments. The method of extracting the crucial subjects from a piece of data to create the patterns or classes in the entire dataset has long been topic modelling methodologies. Despite the similarity in principle, short texts cannot be effectively covered by standard unsupervised topic models; as a result, specialized unsupervised short text topic models were used. These models successfully extract the trending topics from comments and hashtags for additional processing. By utilising the bidirectional processing, these models aid in the extraction of significant issues. However, in order to get sufficient prior information for these unsupervised models, significant training is required, which is not always sufficient. Given these restrictions, a successful strategy for classifying comments and hashtags must be created in order to fill the gap between the classifier and the topic model and greatly improve flexibility.

II. LITERATURE SURVEY

Literature was reviewed from various sources, research papers, these research papers have provided us sufficient amount of data for the survey. The hierarchical approach is followed in the institutional organizations.

In [4]This paper presents a hybrid deep learning model, called DEA-RNN, to detect CB on Twitter social media network. The proposed DEA-RNN model combines Elman type Recurrent Neural Networks (RNN) with an optimized Dolphin Echolocation Algorithm (DEA) for fine tuning the Elman RNN's parameters and reducing training time. They evaluated DEA-RNN thoroughly utilizing a dataset of 10000 tweets and compared its performance to those of state-of-the-art algorithms such as Bi-directional long short-term memory (Bi-LSTM), RNN, SVM, Multinomial Naive Bayes (MNB), Random Forests (RF). The experimental results in this paper show that DEA-RNN was found to be superior in all the scenarios. It outperformed the considered existing approaches in detecting CB on Twitter platform. DEA-RNN was more efficient in scenario 3, where it has achieved an average of 90.45% accuracy, 89.52% precision, 88.98% recall, 89.25% F1-score, and 90.94% specificity.

In [5] Users of online social networks (OSNs) are growing every day, and attacks and threats against users of OSNs have also been growing steadily. Attacks against OSN users take advantage of both system and user-caused weaknesses, which inevitably impact the hacker's attack plan. The objective of this research is to find out how social media users' actions affect how vulnerable they are to security and privacy threats. The study used survey methods and included social media users from Turkey and Iraq. This study records and examines 700 OSN users' actions across two nations. This study analyses the actions of social media users from two different countries to see if there is a correlation between their actions and security and privacy issues. To conclude, this paper analysed social media user behaviours in terms of security and privacy. These paper gives some new knowledge and insights to Security and Privacy Area in terms of user behaviours by considering different kind of security attack scenarios.

In [6] we conduct an extensive survey, covering 1) the multidisciplinary concept of social deception; 2) types of OSD attacks and their unique characteristics compared to other social network attacks and cybercrimes; 3) comprehensive defines mechanisms embracing prevention, detection, and response (or mitigation) against OSD attacks along with their pros and cons; 4) datasets/metrics used for validation and verification; and 5) legal and ethical concerns related to OSD research. Based on this survey, we provide insights into the effectiveness of countermeasures and the lessons learned from the existing literature. This paper

describes various types of OSD attacks in terms of false information, luring and phishing, fake identity, crowd turfing, and human targeted attacks. Following the major OSD types, the comparisons between social network attacks, social deception attacks, and cybercrimes are discussed. And also includes discussed the security breach by OSD attacks based on traditional CIA (confidentiality, integrity, and availability) security goals.

In [7] we present Mal JPEG, a machine learning-based solution for efficient detection of unknown malicious JPEG images. To the best of our knowledge, we are the first to present a machine learning-based solution tailored specifically for the detection of malicious JPEG images. Mal JPEG features are extracted based on the structure of the JPEG image. Mal JPEG features were defined based on an understanding of how attackers use JPEG images in order to launch attacks and how it affects the JPEG file structure in comparison to regular benign JPEG images. The features are simple and relatively easy to extract statically (without actually viewing the image) when parsing the JPEG image file.

In [8] This paper presents a robust methodology to distinguish bullies and aggressors from normal Twitter users by considering text, user, and network-based attributes. Using various state-of-the-art machine learning algorithms, these accounts are classified with over 90% accuracy and AUC. Finally, the current status of Twitter user accounts marked as abusive by our methodology, and study the performance of potential mechanisms that can be used by Twitter to suspend users in the future. The drawback of this paper is the average level performance provided by the state-of-the-art machine learning algorithm and it is susceptible to errors. The paper did not provide real-time detection of abusive behaviors with the aid of properly tuned distributed stream and parallel processing engines. It did not repeat the same analysis on other online social media platforms such as Facebook, Foursquare, and YouTube, in order to understand if the provided methods can detect similar behavioural patterns and can help bootstrap their effort to combat them.

Amanpreet Singh et al. [9] has reviewed many previous research papers related to machine learning models, pre-processing 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

ISSN: 2366-1313

models, pre-processing 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.

III. PROPOSED SYSTEM

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.

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 Pre-processing, Data Extraction will be performed on the fetched Tweets

3. Pre-processed 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 be added to the database and then further 10 tweets from that users' timeline will be fetched, because it cannot 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 pre-processing will be performed over the tweets.

5. 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.

ISSN: 2366-1313

6. 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 pre-processing and feature extraction is performed over the tweets. And after performing pre-processing and feature extraction tweets are passed to the SVM model for classification to predict whether the tweet is Bullying or Non-Bullying.

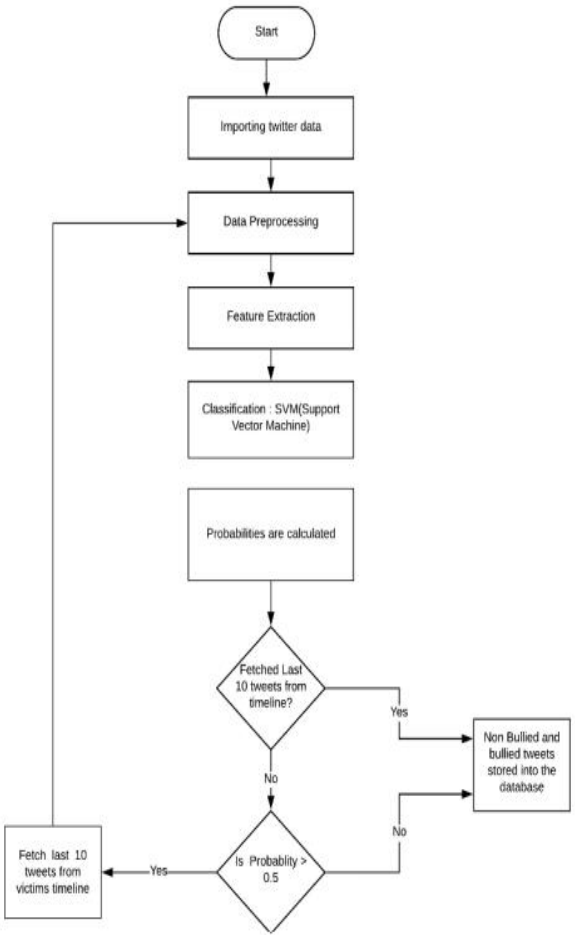

Fig.1 Proposed system architecture

2. Developing the Model:

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

A. PRE PROCESSING:

The Natural Language Toolkit (NLTK) is used for the pre-processing 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

ISSN: 2366-1313

to the list. Firstly, PunktSentence Tokenizer is used to tokenized text into the sentences. 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 pre-processing. 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 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.

IV. RESULTS

In this section, the SVM and Naive Bayes on the dataset collected from the various sources like Kaggle, Github, etc are compared. After performing pre-processing 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

ISSN: 2366-1313

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 7 th is labeled as a "Non-Bullying" tweet.

Table.1 The Accuracy of Support Vector Machine and Naive Bayes

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

Table.2 Classification Report of the Naive Bayes Algorithm

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

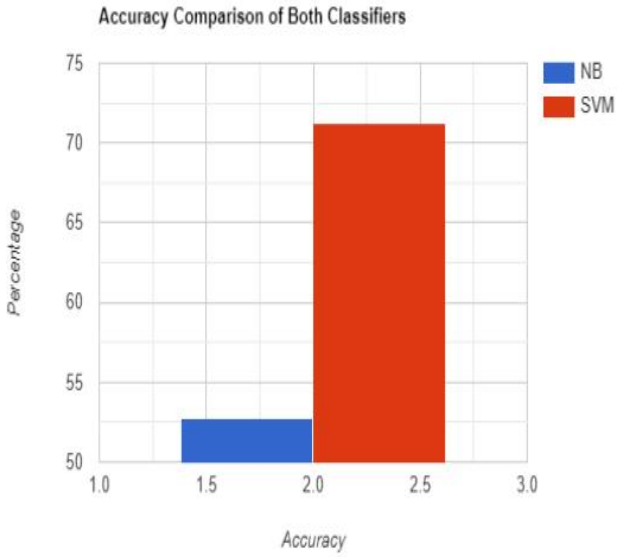

Fig.2 Accuracy Comparison of Both Classifiers

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. F. Mishna, M. Khoury-Kassabri, T. Gadalla, and J. Daciuk, ‘‘Risk factors for involvement in cyber bullying: Victims, bullies and bully–victims’’
2. K. Miller, ‘‘Cyberbullying and its consequences: How cyberbullying is contorting the minds of victims and bullies alike, and the law’s limited available redress’’
3. A. M. Vivolo-Kantor, B. N. Martell, K. M. Holland, and R. Westby, ‘‘A systematic review and content analysis of bullying and cyber-bullying measurement strategies”
4. H. Sampasa-Kanyinga, P. Roumeliotis, and H. Xu, ‘‘Associations between cyberbullying and school bullying victimization and suicidal ideation, plans and attempts among Canadian schoolchildren’’
5. M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong, ‘‘Improving cyberbullying detection with user context’’.
6. G. A. León-Paredes et al., Presumptive Detection of Cyberbullying on Twitter through Natural Language Processing and Machine Learning in the Spanish Language, CHILECON pp. 1-7, doi: 10.1109/CHILECON47746.2019.8987684. (2019)
7. P. K. Roy, A. K. Tripathy, T. K. Das and X. -Z. Gao, A Framework for Hate Speech Detection Using Deep Convolutional Neural Network, in IEEE Access, vol. 8, pp. 204951-204962,, doi: 10.1109/ACCESS.2020.3037073. (2020).
8. S. M. Kargutkar and V. Chitre, A Study of Cyberbullying Detection Using Machine Learning Techniques, ICCMC, pp. 734-739, doi: 10.1109/ICCMC48092.2020.ICCMC-00013 7. (2020)
9. Prasadu Peddi (2015) "A machine learning method intended to predict a student's academic achievement", ISSN: 2366-1313, Vol 1, issue 2, pp:23-37.
10. Prasadu Peddi (2015) "A review of the academic achievement of students utilising large-scale data analysis", ISSN: 2057-5688, Vol 7, Issue 1, pp: 28-35.