Detection of Cyberbullying using Machine Learning

Rupesh Kumar, Shreyas Parakh, C.N.S.Vinoth kumar

Department of Computer Science and Engineering, College of Engineering and Technology
SRM Institute of Science and Technology, Kattankulathur, Chennai
rd2815@srmist.edu.in, sv1588@srmist.edu.in, vinothks1@srmist.edu.in
Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 20 April 2021

Abstract— It's virtually not possible to measures the reasons why cybersecurity is crucial in the expanding battlefield of cybersecurity. Permitting mischievous threats to operate amok anywhere, whenever, and in any circumstance is not acceptable, particularly when it comes to the complex network of client and organization data that cyber security teams are tasked with safeguarding. Individuals and families, as well as corporations, agencies, and educational institutions, must consider cyber security. Machine Learning Can Help Advance the Cyber Security Landscape. Today's businesses collect massive quantities of information/data. Data is considered as the heart of almost every business-focused system imaginable. Infrastructure systems are included in this. In today's era high-tech structure, network and cybersecurity systems, collects massive quantities of data and analytics on most mission-critical systems' key elements. While humans also provide critical serviceable oversight and brilliant insights in today's infrastructure, machine learning and artificial intelligence are rapidly gaining traction in almost all domain of today's systems, be it on premises or on the cloud.

Keywords—cyberbullying, machine learning, naïve bayes, neural networks

I. INTRODUCTION

Cyberbullying is a type of threat that occurs using digital gadgets such as cellphones, hardware systems, and remote tablets. Cyberbullying can occur offline in social media, platforms, or while playing games where people could see, take part in, or share data with one another, or online in social media platforms, or while playing games where people could see, take part in, or share data with one another. Sending, receiving, posting, or sharing negative, dangerous, false or mean content about someone is considered cyber bullying [1]. It may include humiliating or harassing someone by disclosing personal or private data about them. Some cases of cyberbullying may be considered illegal or criminal[2-3].

II. STATE OF THE ART (LITERATURE SURVEY)

There are a number of approaches to recommending applications that can detect online manipulation with remarkable accuracy automatically. The number one author is Nandhini et al., who raised a working-model which refers to the Naive-Bayes machine learning system, and their model received 91 percent accuracy, with their database coming from MySpace.com. They then suggested other model, Nave Bayes organisation and genetic engineering (FuzGen), which also received 87 percent accuracy. Romsaiyud et al. enhanced Nave Bayes classification by extracting words and checking the reunion of a reloaded pattern[4-6]. They got 95.79 percent correctness in data sets from Slashdot, Kongregate, and MySpace using this approach. They also get an issue of interoperable processes that do not function in the same way. Furthermore, Bunchanan et al. use the game chats from War of Tanks to retrieve the dataset again, isolate it by hand, or compare it to easy ones in this method. As compared to the results divided by hand, the Nave group, which uses emotional analysis as a factor, had bad results. In addition, after obtaining their database from [7] kaggle, Isa et al. suggested using two splits: Nave Bayes and SVM. The Nave Bayes divide has a 92.81 percent average accuracy, while SVM has a poly kernel revealed 97.11 percent accuracy, but they didn't define their own practice or test size of database, so the outcomes mightn't be accurate. An alternative to Dinakar et alquest .'s for explicit and explicit language about (1) sexual acts, (2) race and culture, and (3) wisdom is to build their own database from YouTube comments. After implementing SVM and Nave Bayes preparation, SVM posted a 66% accuracy rate and a 63% Nave Bayes rate. Continuing on from [8-10] Di Capua et al., see proposes a new way to find web-based identification by using an unregulated tool, using separators inconsistently in addition to their database, using SVM in FormSpring to enforce 67 percent remembering, using GHSOM on YouTube to achieve 60 percent accuracy, 69 percent accuracy, and 94 percent memory, and using Nave Bayes on Twitter to achieve 67 percent recollection. Furthermore, Haidar etal. introduced a working-model for detecting cyber bullying, rather they used the Nave Bayes and got 90.85 percent accuracy, while SVM got 94.1 percent accuracy, but they got a comparatively high degree of wrong detection and operations in Arabic. that considers your paper to be a component of the entire proceedings rather than a stand-alone document Please don't change any of the existing titles[11-13].

III. PROPOSED WORK

Preprocessing, feature extraction, and classification are all stages in the classification process. By removing the outliers, noise we clean the data in the preprocessing phase Pre-processing procedures:

- Making tokens: We take text as sentences or full parts in this section and extract the inserted text as words from the list.

- Downgrade text: This lowers all the letters in a list of found words without the token, so 'THIS IS AMAZING' becomes 'this is amazing'.

- Set words and code coding: This is a vital step in the process since we clean up the text by setting encoded properties like '\n' or '\t' that does not provide valuable subject matter to distinguish.

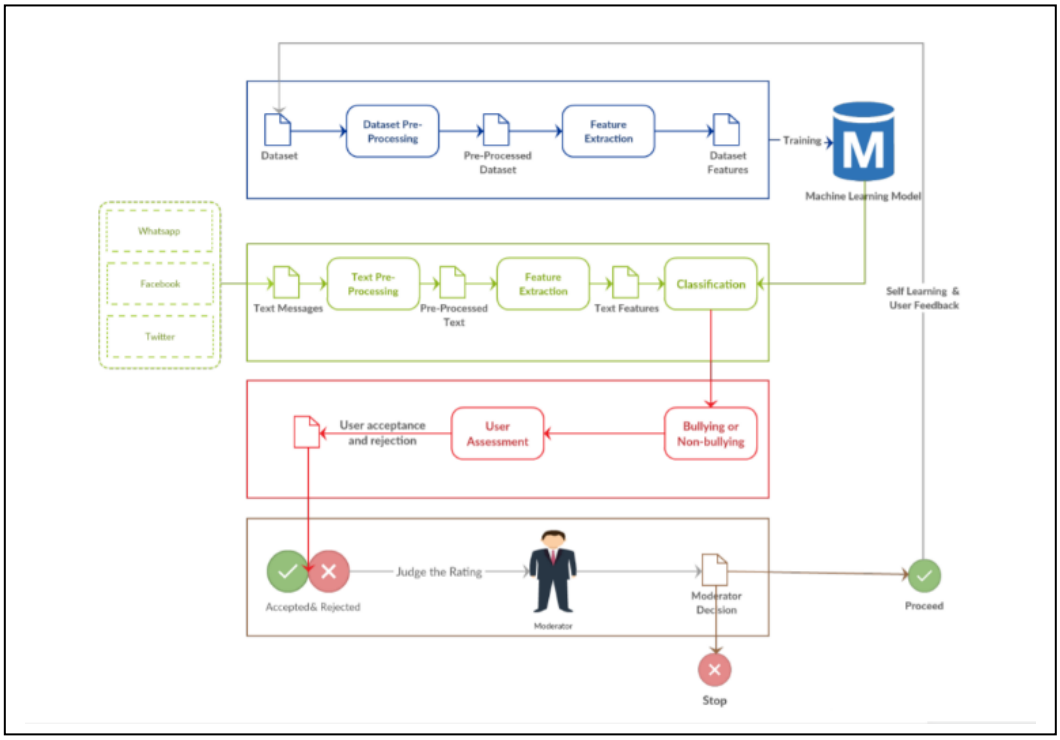

Fig. 1- Voice rectification

In this , we will use Microsoft Bing's word correction API, which takes input of a word and replaces it with a JSON item with very related names based on the indifference between these words and the first voice.

The projected Model's second step is to remove functionality. This phase converts the text data into a format that can be used to feed machine learning with high performance. First we extract the input information features using TFIDF as it then puts it on the list of factors. The main idea of TFIDF that it works on text and weighs words related to a sentence. In addition in TFIDF, we used the emotional analysis process to extract sentence size and include it as a feature in the file a list of features containing TFIDF features. Unity of sentences mean that if the sentence is defined as correct or negative. To that end, we use Post library Blob, a pre-exercised/trained model in movie reviews, to delete the fakes. The lifting technique uses N-Gram to deal with various combinations of words during model testing, in addition to function removal using TFIDF and emotional release. In general, we use 2-Gram, 3-Gram, and 4-Gram[14-16].

The final step in this proposed approach is categorization the step at which the extracted elements are fed in phases algorithm for training, and testing for division and apply it to forecast phase. SVM (Vector Support Machine) and Neural Network are the two types of partitions we used. Nervous system The network is made up of three layers: input, hidden, and output[17]. There are 128 nodes in the input layer. It has 6 neurons and is located

in a secret layer. The Boolean release is the output layer. Typically, classifiers tests are performed using several test matrices depend on the matrix of confusion. In between of those methods accuracy, precision, recall and f-score.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

$$F-Score = \frac{2*precision*recall}{precision+recall}$$

Calculations are done by considering the following statistics:

TP - true positive value, TN - number of true objections, FP - false value, and FN - value of classes of negatives.

A. Abbreviations and Acronyms

NBC-	Naive Bayes Classifier	
SVM-	Support Vector Machine	
AI	-	Artificial Intelligence
ML	-	Machine Learning
TFIDF	-	Term Frequency-Inverse Document Frequency

IV. IMPLEMENTATION

The experimental outcomes of the proposed method are elaborated in this part. On the cyberbullying dataset from Twitter, we test the proposed method. There are 1065 datasets in total. The Data and the Findings are described in the following sections.

We used a Twitter cyberbullying dataset that had been compiled and labelled. In general, this dataset includes a large number of Twitter communication posts. The dataset includes questions and responses that are either annotated with cyberbullying or not.We have used two files to run the code. First one is used for twitter API call and the second one to implement algorithms, to train and test data set. We've divided dataset into the ratio as shown in the fig.

```
#Split the data into training and test
train_set, test_set = Final_Data[0:746], Final_Data[746:]
```

For text extraction from the Twitter API we've used NLTK library. Model is evaluated with n-grams(unigrams, bigrams, trigrams, combination (n=3). We've classified bullying and non-bullying activities by labeling them as 1=Bullying and 0=Non-Bullying.

There are many algorithms and libraries used in this project -

1. Naive Bayes
2. Support Vector Machine

V. RESULTS DISCUSSION

We repeat the preprocessing phase to extract the features after preprocessing the dataset. The dataset was then divided into two parts: train and test. To test the classifiers, accuracy, recall, and precision, as well as the f-score, are used as performance measures. We use Naive Bayes and SVM because they are among the best performing classifiers available.

	Tweet	Text Label
0	.omg why are poc wearing fugly blue contacts s...	Non-Bullying
1	.Sorry but most of the runners popular right n...	Non-Bullying
2	.those jeans are hideous, and I?m afraid he?s ...	Non-Bullying
3	.I had to dress up for a presentation in class...	Non-Bullying
4	.Am I the only one who thinks justin bieber is...	Non-Bullying
...	...	...
1060	No we are not, But you are a race baiting libt...	Bullying
1061	you wont get anyone for this challenge., after...	Bullying
1062	I will follow you if you are not a libtard,Mus...	Bullying
1063	michaelianblack Ur a child, an ostrich w/ your...	Bullying
1064	FoxNews. not to all the ppl I know that live t...	Bullying

```
Naive Bayes Performance with Unigrams
Accuracy: 0.6614420062695925
UnigramNB Recall
Bullying recall: 0.5962732919254659
```

Figure 2 Showing the Text label of Cyber bullying content

Figure 3Performance accuracy of

```
UnigramSVM Recall
Bullying recall: 0.7021276595744681
1065
Naive Bayes Performance with Bigrams
Accuracy: 0.6572327044025157
```

Naïve Bayes

We performed multiple studies using various n-gram language models. At the time of estimation of the model generated by the classifiers, we pay special attention to 2-gram, 3-gram, and 4-gram.

```
Naive Bayes Performance with Trigrams
Accuracy: 0.5880503144654088
bullying precision: 0.7857142857142857
bullying recall: 0.07913669064748201
```

Figure 4 Performance accuracy of Naïve Bayes by recall

Figure 5 Performance accuracy of Naïve Bayes with Trigrams

VI. CONCLUSION

We suggest a framework for detecting cyberbullying using machine learning methods in the paper. Naïve Bayes classifier and SVM are used to evaluate the model. And after a lot of practise, we've decided that SVM with N-

grams is the best way to go. Our work can improve cyberbullying identification and help people use social media safely by achieving this level of accuracy. The size of the training data, however, limits the detection of cyberbullying patterns. We will benefit from having access to a large dataset. As a result, deep learning techniques may be appropriate because they have been shown to outperform machine learning approaches over large datasets.

VII. REFERENCES

[1] John Hani, Mohamed Nashaat, Mostafa Ahmed, Zeyad Emad, Eslam Amer, "Social Media Cyberbullying Detection using Machine Learning", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 5, 2019, Pages 703–707, Published- 04 February 2019

[2] Sinchana, Sinchana, Pradyumna C, Janhavi, Deepika, "Detection of Cyberbullying using Machine Learning", International Journal for Research in Applied Science & Engineering Technology (IJRASET). Volume 8 Issue VII July 2020, Published - 7 July 2007

[3] Ghada M. Abaido , "Cyberbullying on social media platforms among university students in the United Arab Emirates", International Journal of Adolscense and Youth, Published- 14 September 2019

[4] B.Sri Nandhinia, J.I.Sheebab, "Online Social Network Bullying Detection Using Intelligence Techniques", Elsevier B.V.

[5] 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.

[6] Shane Murnion, William J Buchanan, Adrian Smales, and Gordon Russell. Machine learning and semantic analysis of in-game chat for cyberbullying. Computers & Security, 76:197–213, 2018.

[7] 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.

[8] A.Saranya, R.Naresh "Cloud Based Efficient Authentication for Mobile Payments using Key Distribution Method", Journal of Ambient Intelligence and Humanized Computing, Springer, 02 January, 2021. https://link.springer.com/article/10.1007%2Fs12652-020-02765-7

[9] R.Naresh, P.Vijayakumar, L. Jegatha Deborah, R. Sivakumar, "A Novel Trust Model for Secure Group Communication in Distributed Computing", Special Issue for Security and Privacy in Cloud Computing, Journal of Organizational and End User Computing, IGI Global, Vol.32, No. 3, Septemer 2020, Pp. 1-14. DOI: 10.4018/JOEUC.2020070101

[10] R.Naresh, M.Sayeekumar, G.M.Karthick, P.Supraja, "Attribute-based hierarchical file encryption for efficient retrieval of files by DV index tree from cloud using crossover genetic algorithm",Soft Computing, Springer, Vol.23, No. 8, 2019, Pp. 2561-2574. [Impact Factor=3.050]

[11] R Divya Mounika, R.Naresh, "The concept of Privacy and Standardization of Microservice Architectures in cloud computing", European Journal of Molecular & Clinical Medicine, Vol 7, No 2, Pages 5349-5370, Dec 2020.

[12] P.Vijayakumar, R.Naresh, L. Jegatha Deborah, SK Hafizul Islam, "An efficient group key agreement protocol for secure P2P communication", Security and Communication Networks, Wiley, Vol.9, No.17, pp.3952–3965, 2016 http://onlinelibrary.wiley.com/doi/10.1002/sec.1578/abstract

[13] P.Vijayakumar, R.Naresh, SK Hafizul Islam, L. Jegatha Deborah "An Effective Key Distribution for Secure Internet Pay-TV using Access Key Hierarchies", Security and Communication Networks, Wiley, Vol.9, No.18, pp.5085–5097, 2016.

[14] R. Naresh, M Meenakshi, G Niranjana, "Efficient study of Smart Garbage Collection for Ecofriendly Environment", Journal of Green Engineering, Vol.10, No.1, pp.1-10, Feb 2020.

[15] Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. ACM Transactions on Interactive Intelligent Systems (TiiS), 2(3):18, 2012.

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

[17] Batoul Haidar, Maroun Chamoun, and Ahmed Serhrouchni. A multilin gual system for cyberbullying detection: Arabic content detection using machine learning. Advances in Science, Technology and Engineering Systems Journal, 2(6):275–284, 2017.