A STUDY OF HATE SPEECH DETECTION USING DIFFERENT MODELS

Mihir.Y, Koushik.K, Mahadev.C, Sujith S
Department of Computer Science & Engineering, GITAM School of Technology, GITAM University, Visakhapatnam

ABSTRACT - This paper focuses on a comparative study of algorithms that can detect social hate speech/mongering on the cyber world. The act of using online platforms to harass or bully others, known as cyberbullying, is a significant problem that has serious consequences for its victims. Hate Speech is one such issue under this category. Hate speech leads to violence, bullying, harassment and disturbances. One of the primary forms is text on social networks, affecting over billion users. Despite the use of machine learning models, including deep learning, to tackle cyberbullying, the challenge of effectively classifying and monitoring such behaviour still persists widely. To tackle this issue, an existing approach involves the use of a novel concept of CNN. The proposed approach is to understand different machine learning model such as Logistic Regression, Naïve Bayes, Decision Tree, Random Forest and BERT. The use of NLP techniques while training the model can improve its performance in processing and analysing textual data, the model can better understand the language used in the dataset and accurately classify instances of cyberbullying. The goal is to effectively monitor and prevent cyberbullying conducted in the form of Hate speech text, using machine learning and language processing techniques.

Index Terms—Hate speech , Machine Learning, NLP Approach.

I. INTRODUCTION

Modern society is increasingly dependent on online apps and social media platforms for e-commerce, digital payments, services and social networking. The "Technology" mammoth is growing uncontrollably. It has reached its great heights where nothing is impossible yet equally capable to transform into a ruthless monster. The ill impacts of the technology are implemented in many forms, embodiments or avatars. One such manifestation is via Social networking and the devil in disguise is Cyberbullying. It is served on a platter full of rosy pictures of making you connected, influenced, popular and worthy[1]. The Social media attracts youth, children and olds alike with their customized emptiness. Cyberbullying is the cruel act of targeting and using the vulnerable set of people. It does not discriminate, caste no bar, age no bar, it claws out at everyone. The COVID times has also contributed to worsen the Cyberbullying issue. Due to the shutdown and lockdown of face-on communication, the need or compulsion for online communication is increased. Cyberbullying can be as small as messaging gibberish, sending fake promotions or as big as sending sexual junk email etc[2].

Use of different social media such as Twitter, Instagram, Facebook, etc. is increased multi-fold. These and many such sites are where cyberbullying or Hate speech have been reported. Exploit of social media has made hate speech escalate to a zenith. Many countries lack a clear legal framework to tackle cyberbullying, posing challenges to effectively addressing the problem[3] The hate speech is defined as derogatory words or statements on the platform. Different researches show that more than 80 per cent of people/youth are exposed to hate speech on the receiving end[4]. The anonymity of the offender further complicates the issue, leading to its increased interest in research.. Nowadays, we can see a rise in the amount of this wretched curse and it is mostly directed in attacking teenagers and certain highly vulnerable set of people. It is also targeted to the most influential people for obvious reasons. Hate speech might turn into either depression or suicidal attempts if impacted the individuals[5].

It is important to have a reliable system in place to identify instances of cyberbullying and provide assistance to affected individuals. Cyberbullying can occur across multiple social media platforms, which makes it difficult to track and address. Organizations are trying to incorporate policies and terms to minimize such incidents. eg. Twitter has implemented community guidelines to address issues such as online bullying, abusive behaviour, sexual assault, violence against public figures, and illegal activities. Similarly, Facebook and other platforms also have guidelines to support victims of social abuse. To protect adolescents from cyberbullying, it is essential to detect instances of text-based cyberbullying, the most common form of online harassment used by perpetrators. As this is quite on the rise among others offences.

There is a growing interest in computer science to develop automated systems that can detect and mitigate episodes of cyberbullying. In this project study, we propose different approaches to tackle hate speech on different social media platforms and on the internet. Existing systems are equipped to identify bullying content. Hence, this study aims to focus on identifying text-based cyberbullying. The aim of such algorithms is to understand the statement or the sentence

from the given dataset and find out the output[6]. There are different machine learning approaches for tackling the problem, the existing approach uses the CNN algorithm to train and test the model. This research is about selecting the correct algorithm for a certain dataset. Our approach focuses on NLP and Machine learning algorithms, such as Logistic Regression, Naïve Bayes, Decision Tree , Random Forest and BERT compare all algorithms to find the best of those for a certain dataset[7]. This detailed study will contain graphs and tables to compare and demonstrate the precision and accuracy obtained for different models.

II. LITERATURE SURVEY

Nowadays we have witnessed immense growth in social networking which provides a playground for conflict which may be sexism, racism or religious disparities. This leads to hate speech which is the use of unparliamentary language and rude content. They probed into various sites like Twitter to present their outcomes. Their strategy relies on using individual words (unigrams) and patterns as the fundamental building blocks. These building blocks are then utilized as features to train machine learning models, and they are extracted from the training data[8].

But with the rapid growth of new words and unknown words the hate mongering haven't been lessen. So this research method uses the way words are pronounced as characteristics to identify prohibited terms and also assists in detecting duplicate words, leading to a reduction in the number of attributes. The creators were influenced by the Bag-of-Words technique, which is extensively used for extracting textual features[9]. So with the improvement of technology and reach of social media to everyone caused a havoc on the environment. Different researchers tried to mitigate the cause and help for the future improvements. The primary aim of this research is to conduct a comprehensive evaluation of existing text-based systems that detect hate speech. The evaluation includes an analysis of the datasets, textual features, and machine learning models utilized in these systems. Moreover, more than 100 relevant papers were subjected to content analysis to identify common themes. This review's results provide an organized and informative summary of the current state of identifying hate speech in textual data[10].

The Machine learning concepts weren't that upto the mark for some researchers, whole accuracy depends on how the dataset is and what are the variables in the dataset. Some used ANN and LSTM. A proposed system utilizes Twitter hashtags to identify tweets as either hate speech or not. To accomplish this, the system uses the LSTM technique as a classifier. Each tweet that is entered into the system receives a label of "hate speech" or "non-hate speech"[11].

III. PROBLEM STATEMENT

Social media platforms such as Facebook and Twitter are particularly notorious for being hotspots for cyberbullying, and this is a significant problem in modern times. This causes havoc on the environment and on the youth's life. Gloominess on the face of every youth, the sadness of failing in every circumstance and discouragement from bullying forms an inferiority complex and make intrinsic thoughts damage the ferocity of the youth. Words cut deeper than a knife and these bullies are very well aware of that. They target people and aim for topics that are person specific and sensitive. This whole race, leaves behind humanity and any traces of compassion, kindness and empathy. They make believe that they are feared monsters and gnaw on the received attention in the form of fear. This makes them forget that we are all humans like they were once and we all have problems like them.

Identifying and halting cyberbullying is a complex task as it typically takes place online and is frequently concealed from the watchful eyes of educators and parents. To address this issue, new strategies are needed to automatically detect and prevent instances of bullying on social media platforms. In this regard, researchers have investigated a system that is capable of autonomously identifying and reporting cases of cyberbullying on social media platforms.

This research aims to devise strategies to combat cyberbullying and hate speech and enhance the internet's condition by analyzing the extent of the negative impact of hate speech on individuals. In order to accomplish the objective of developing an automated system for detecting cyberbullying on social media platforms, extensive research is conducted to gain a deeper understanding of the nature and extent of this problem. This is followed by a review of previous studies on detecting and preventing bullying in order to determine the most effective algorithms and techniques. Based on the findings of this research, a suitable system architecture is designed and implemented, which includes a model capable of real-time identification of instances of bullying on a chosen social media platform.

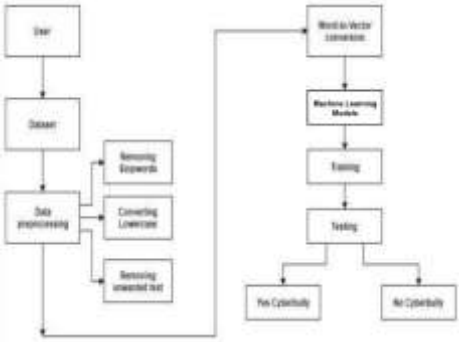
A. EXISTING SYSTEM

The current system faces challenges in implementing machine learning algorithms due to limited information about data visualization. Moreover, the existing system relies on mathematical calculations for building models, which can be time-consuming and complex. To address these issues, we propose to use machine learning packages from the sci-kit-learn library and different machine learning models which can simplify the model-building process and provide more efficient solutions.

B. PROPOSED SYSTEM

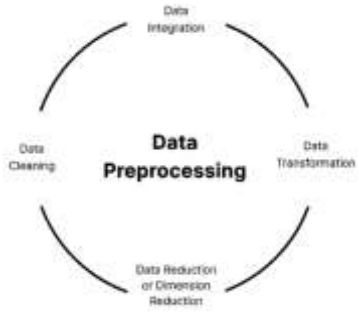
Numerous machine learning models have been proposed for classifying hate speech. However, none of them have successfully addressed the issue of misdiagnosis. Furthermore, prior research on performance evaluation for hate speech classification models typically disregards data heterogeneity and size. Thus, we suggest utilizing Natural

Language, as well as Logistic Regression, Naïve Bayes, Decision Tree Random Forest, LSTM and BERT techniques, to determine whether a given speech constitutes hate speech or not. The research will also show which algorithm is best for the hate speech classification problem on a certain dataset. The advantages of the proposed systems are more accuracy and precision and reduces time complexity.



As the figure shows the steps of the given project go as – 1) Searching for the Dataset, it is found at different websites on the internet such as Kaggle, Google Dataset, Datahub.io and UCI Machine Learning Repository. Select and Understand the dataset with which you want to proceed. Our proceed systems, showing of different Machine Learning implementation is included.

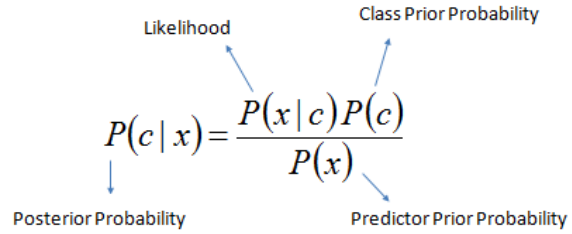
2) The preprocessing of data is a vital step in machine learning, where raw data is transformed into a format that is more appropriate for training a model. This procedure is crucial since it enables the data to become more manageable and interpretable, which in turn helps the machine learning model to learn effectively and produce precise predictions. Apart from cleaning the data by removing irrelevant words and special characters, this step may also require combining data from various sources, manipulating the data, handling missing values, selecting important features, encoding categorical variables, and dividing the data into training and testing sets. The accuracy and effectiveness of a machine learning model depend heavily on how well the data has been pre-processed. It includes important steps such as – a) Data Integration b) Data Transformation c) Handling of missing data from the dataset d) Splitting the data and etc. The purpose of data pre-processing is to convert raw data into a suitable format that can be used to train a machine learning model. The aim is to ensure that the data is presented in a format that can be easily analysed and interpreted by machine learning algorithms so that accurate predictions or classifications can be made. For example suppose your raw data is "Why is it that foreigners smell weird? Do they.", after data pre-processing the data which it reads is "why is it that foreigners smell weird do they ..." The step remove the capitalized letters and unwanted words such as any special characters like { ‘ ; . \ > @ ? [ ] } etc which are basically not required for the training of the model, and also reduces the size of the data set which has to train. The models only accept the vector format so the data pre-processing steps make sure that the sentence is pre-processed and changed into the vector format[12].



3) Models/ Algorithms used in this research are – Machine Learning models which are best for classification problems. The hate speech recognition shows if it is Hate speech or Not Hate speech. It is a binary classification problem so the models used are logistic regression, naïve Bayes, Decision Tree, Random Forest and LSTM.

Logistic Regression - Logistic regression is a machine learning algorithm that is frequently used for binary classification problems, where the outcome involves only two possible class values. For example, if the student will pass the exam or not (“YES” , “NO”) by seeing the time he spent studying. In this project Logistic Regression is shown as if it is hate speech or not hate speech.

Naive Bayes – For the classification task Naïve Bayes is used. It works by applying the Bayes theorem to find the probability of a hypothesis given some evidence. The classifier assumes that the occurrence of one feature does not affect the occurrence of another. It then calculates the probability of each possible output class based on the input features and assigns the class with the highest probability as the predicted output. The Naive Bayes classifier is widely used in areas such as text classification, spam filtering, sentiment analysis, and recommendation systems.

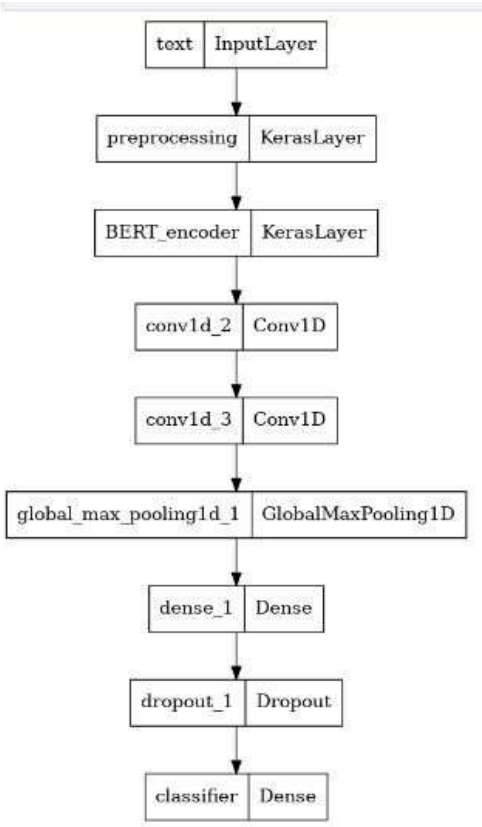


$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Decision Tree- The supervised learning technique is commonly utilized in Machine Learning to address classification problems. The decision rules are represented by the branches, the outcomes are represented by the leaf nodes, and the features of the current dataset are represented by the internal nodes. For this project the Decision tree is used since it is also a classification problem (Binary Classification with "Yes" and "No") and decision tree is one of the used models for either classification or a regression problems.

Random Forest – Random forest is an upgraded and modernized variant of the decision tree algorithm. It differs from a single decision tree by utilizing multiple decision trees that collaborate to provide more accurate and feasible results. As a result of this collective approach, it outperforms the decision tree in terms of accuracy and feasibility.

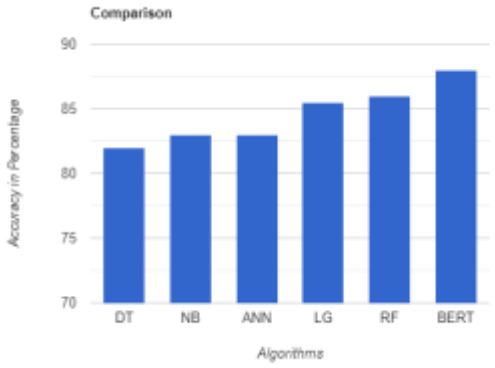
BERT– Transform encoders have achieved significant success in natural language processing (NLP). BERT is one such encoder that has gained wide usage in deep learning due to its ability to provide vector-space representation. By utilizing transformer architecture, BERT processes each input token efficiently. BERT stands for Bidirectional Encoder Representation from Transformation, and the Small BERT model is a base model that features 12 encoders stacked on top of each other. Due to its architecture, Small BERT performs faster and is more suitable for binary classification tasks and small datasets[13].



1 Comparison and Results

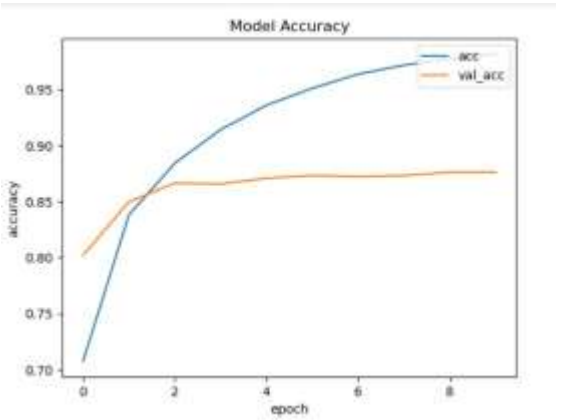
ACCURACY OF DIFFERENT MODELS

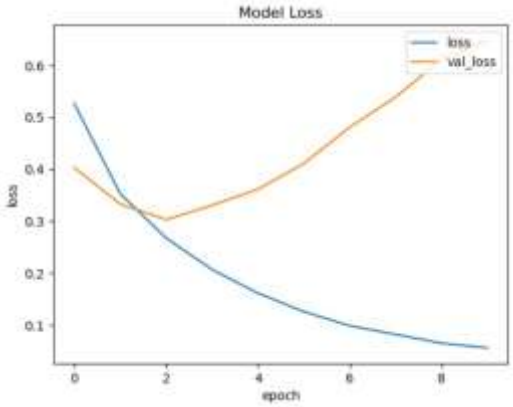
S.No	MODEL NAME	ACCURACY OBTAINED
1	BERT	0.885
2	Random Forest	0.858
3	Logistic Regression	0.857
4	Naive Bayes	0.834
5	Decision Tree	0.828



In figure – DT:- DECISION TREE, NB:- NAÏVE BAYESS, ANN, LG:- LOGISTIC REGRESSION, RF:- RANDOM FOREST, BERT

From the current dataset, we noticed that random forest has been the top-performing machine learning model. However, when compared to BERT, it fell short. BERT exhibited superior accuracy and greater suitability for the task. If we go through the accuracy BERT has a top-notched performance, but machine learning models have a better advantage while comparison to time complexity. Machine Learning models work with less time complexity than BERT.





BERT(Bidirectional Encoder Representations from Transformers) is a deep-learning model which performs better in comparison to other models. Hate-mongering detection requires high accuracy, precision and less time complexity.

BERT Vs LSTM

Model Name	Loss	Avg Epoch Time	Accuracy	Total Epoch Run
LSTM	0.55	1200	0.77	3
BERT	0.26	220	0.86	3

Compared to LSTM and machine learning models, BERT performs better. BERT achieves higher accuracy than machine learning models and has lower time complexity than the LSTM model. So Small BERT might be a perfect option for Hate mongering detection for a current dataset.

IV. CONCLUSION AND FUTURE SCOPE

It's important to acknowledge that every model has its limitations. Our current models only support binary classification, meaning they can only classify inputs as either "Yes" or "No." However, we aim to enhance our research by incorporating a new label for identifying abusive language. We compared BERT with other machine learning models and other deep learning models. Our research shows that BERT models shows increase in accuracy when compared to other machine learning models and is more suitable for hate mongering detection.

V. REFERENCES

[1] Areej Al-Hassan and Hmood Al-Dossari: Detection of Hate Speech in Social Network: A Survey On Multilingual Corpus. 6th International Conference on Computer Science and Information Technology.

[2] Badjatiya, P., Gupta, S., Gupta, M., Varma, V.: Deep learning for hate speech detection in tweets. In: Proceedings of the 26th International Conference on World Wide Web Companion. pp. 759–760 (2017)

[3] Patchin, J.W., Hinduja, S.: Bullies move beyond the schoolyard: A preliminary look at cyberbullying. Youth violence and juvenile justice 4(2), 148–169 (2006)

[4] Saloni Mahesh, Vdiya Chitre: Study of Cyberbullying Detection using Machine Learning Techniques. 2020 Fourth International Conference on Computing Methodologies and Communication

[5] Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., Bhamidipati, N.: Hate speech detection with comment embeddings. In: Proceedings of the 24th international conference on world wide web. pp. 29–30 (2015)

[6] Nugroho , E Noersasongko , Purwanto, Muljono, Ahmad Zainul, Affandy, Ruri Basuki: Improving Random Forest Method to detect hatespeech. 2019 International Conference on Information and Communications Technology

[7] Fariha Tahosin Boishakhi ,Ponkoj Chandra Shill, Md. Golam Rabiul Alam: Multi-modal Hate Speech Detection using Machine Learning : 2021 IEEE International Conference on Big Data (Big Data)

[8] M. Bouazizi, H. Watanabe and T. Ohtsuki, "Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection," in IEEE Access, vol. 6.

[9] A. Shekhar and M. Venkatesan, "A Bag-of-Phonetic-Codes Model for Cyber-Bullying Detection in Twitter," 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)

[10] M. Gomes, R. Martins, J. J. Almeida, P. Henriques and P. Novais, "Hate Speech Classification in Social media Using Emotional Analysis," 2018 7th Brazilian Conference on Intelligent Systems (BRACIS)

[11] S. S. Syam, B. Irawan and C. Setianingsih, "Hate Speech Detection on Twitter Using Long Short-Term Memory (LSTM) Method," 2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)

[12] Usman Naseem, Imran Razzak & Peter W. Eklund :A survey of pre-processing techniques to improve short-text quality: a case study on hate speech detection on twitter : Published: 04 November 2020

[13] Khouloud Mnassri, Praboda Rajapaksha, Reza Farahbakhsh, Noel Crespi: BERT-based Ensemble Approaches for Hate Speech Detection