COMPARATIVE ANALYSIS OF DIFFERENT MACHINE LEARNING METHODS FOR HATE SPEECH RECOGNITION FOR TWITTER TEXT DATA

Bishal Mohanty
Department of Computer Science and Engineering
Silicon Institute of Technology
Bhubaneswar, India
cse.20bcsb98@silicon.ac.in

Subashis Sahoo
Department of Computer Science and Engineering
Silicon Institute of Technology
Bhubaneswar, India
cse.21bcsl07@silicon.ac.in

***Abstract*—The objective of this paper is to address the problem of detecting hate speech on social media platforms by developing a more accurate machine learning model. With the widespread use of the internet and mobile phones, hate speech has become a prevalent problem in online communication. The research team utilized the well-known Davidson dataset ,which is specifically designed for hate speech recognition on Twitter. They employed different machine learning algorithms and compared their performance based on various parameters, including accuracy, precision score, recall, and F1 score. Through their analysis, they determined that the XGBoost algorithm, when coupled with the TF-IDF transformer embedding technique, achieved the highest accuracy rate of 94.43**

***Index Terms*—Hate speech recognition, Machine learning, Social media platforms, XGBoost algorithm**

I. Introduction

The study of human activities has always been an active research area, and with the fast growing of digitalization and social networking sites, people have become closer to one another, regardless of their diversity and backgrounds. The result of this has been an increase in digital conflicts between individuals, groups, and even communities. This is making hate speech the hottest topic in the field of Machine Learning for researchers and governments. To keep their platforms safe for users, social media companies are investing a significant amount of money in filtering negative, hateful content.

But, there is no clear fixed scale to differentiate between hateful and offensive speech, nor is there a formal definition to differentiate the two. But for the sake of simplicity, we can define any group of words or sentences that hurt the feelings of any person or group and cause them harm emotionally, as hate speech. Most places have defined certain rules and regulations based on their beliefs, and everyone should act within the rules. These rules and laws also extend to social media platforms, and even those platforms have provided certain rules to prevent hate speech. Violating any of the rules may cause significant fines or even imprisonment.

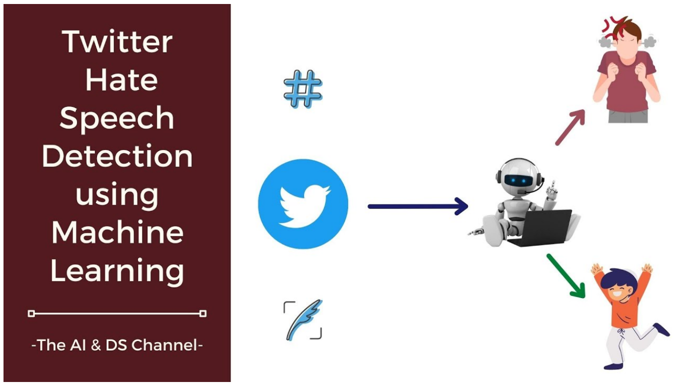


Fig. 1. Hate Speech Recognition

Despite these efforts, Twitter and Facebook have faced significant criticism for not doing enough to ensure that the feelings of their users and communities are not hurt by any means. They are even forced to face the laws related to common people and their differences, such as race, color, gender, sexual orientation, and more. Based on the above discussion, hate speech is defined as any language or gestures in physical form that can deliver hatred toward a targeted group of people to humiliate and insult its members. In extreme scenarios, it can even lead to violence, but for the sake of this paper, we will restrict our discussion only to hate speech.

It should be noted that cultural diversity should also be taken into account because something offensive to one community may not be considered offensive in another. For example, the word teenagers use in everyday conversations may not be suitable for posting on Facebook or Twitter. This term was considered in most of the previous works, but there was still a lot of confusion and miscommunication.

Due to such vast networks, it is impossible to detect and classify each and every piece of content on the internet manually. Hence, we need automated methods to detect hate

Identify applicable funding agency here. If none, delete this.

speech, and in this paper, we have used the Davidson dataset [10]. Each tweet in this dataset is manually classified as “hate,” “offensive,” or “neither” by the author, making it an accurate dataset to work with.

The main player and opponent in the way of automatic hate speech detection on social media and other communication-related platforms is to separate hate speech from other families of impressions and wordings nearly belonging to the same categories and hierarchies. Hence, we implemented two algorithms named Logistic Regression and Naïve Bayes. The latter model outperformed the other model and gave the best results among the model.

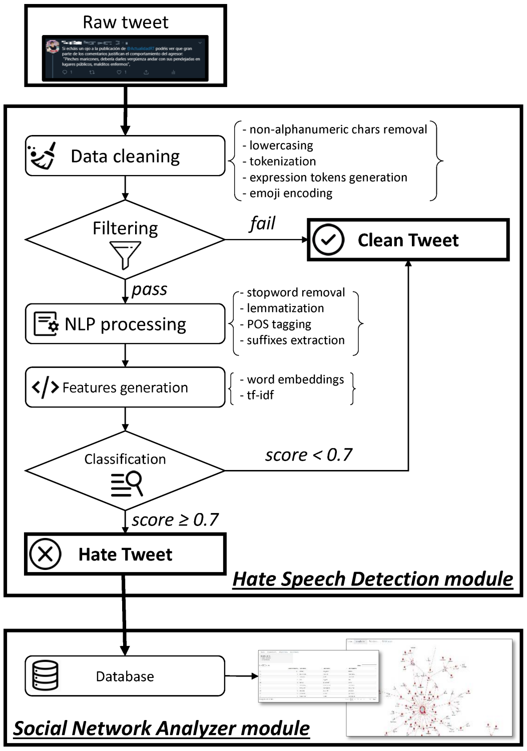
In the next part of this paper, we will discuss the related work that has been done in this field by many authors, followed by the details of the dataset we have used in this paper. We will delve into the intricacies of these algorithms and analyze the results they provide. In section four, we will discuss the models we implemented, followed by the results obtained by these models and a comparative study of these models on different parameters. Finally, in the last part of the paper, we will conclude our work with some possible future scope in this field.

Overall, the study of hate speech detection on social media platforms is a vital area of research. With the rise of digital communication, it has become increasingly important to ensure that online platforms are safe and inclusive for everyone. Automated methods like those discussed in this paper are essential

II. Related Work

Detecting hate speech is a critical challenge for social media platforms such as Facebook, Twitter, and Reddit [1]. To address this challenge, these companies have invested massive amounts of resources into creating systems and tools to monitor and filter content that is deemed hateful or offensive. For instance, Twitter employs a real-time mechanism to filter and monitor hate speech. They have developed a dashboard that continually scrutinizes every tweet, using an algorithm that flags or takes down tweets based on their severity [2]. Similarly, Facebook uses an advanced filtering system that incorporates machine learning algorithms to detect and classify hate speech in real-time. These systems have proven to be highly effective in identifying and removing harmful content [3].

While most hate speech detection systems are built for English language content, there is growing interest in developing multilingual models for other languages, including Hindi, French, Spanish, Arabic, and more [4]. Many researchers are actively working on creating robust and accurate models to detect hate speech, using cutting-edge deep learning methods, including Recurrent Neural Networks (RNNs). Despite significant progress, this field is still advancing, and researchers are continuously seeking to improve the accuracy and effectiveness of hate speech detection systems. The need for such systems has never been more crucial, given the ever-increasing use of social media and the growing concern over the impact of hate speech on individuals and society.



Si echáis un ojo a la publicación de @ActualidadRT podéis ver que gran parte de los comentarios justifican el comportamiento del agresor: "Pinches maricones, debería darles vergüenza andar con sus pendejadas en lugares públicos, malditos enfermos",

Fig. 2. Sequence of Processes

III. Dataset

The hate-speech-and-offensive-language dataset, also known as the Davidson dataset [10], is a vast compilation of words, phrases, and sentences that have been labeled as hate speech by internet users, and it was created by Hatebase.org. The dataset was constructed by gathering a large sample of tweets from 33,458 users using the Twitter API, resulting in 86 million tweets, which were then narrowed down to a sample of 25,000 tweets. These tweets were then categorized into three groups: ”0” for hate speech, ”1” for offensive language, and ”2” for neither. For the purpose of this study, only two classes were used, namely hate and non-hate, with offensive language being considered as hate. The categorization was done by using CrowdFlower workers who were paid for their services. The tweets were cross-checked while categorizing to ensure that they were accurately labeled. While the Waseem dataset [11] also attempted a similar task, it only contained 15,000 tweets, which was deemed insufficient for these models. Therefore, the Davidson dataset was used instead. The dataset is in CSV format and is 4 MB in size. The dataset includes several crucial columns, including the count column, which indicates the number of people who analyzed a tweet, the hate speech column, which indicates the number of users who consider a tweet as hate speech according to its basic definition, the offensive language column, which shows the number of users who consider a tweet as offensive speech, and the neither column, which shows the number of users who consider a tweet as neither hate nor offensive speech. The class column contains the class number of the group that received the maximum votes, while the tweet column contains the actual tweet text.

1) **Column Count:** The count is the number of people who analyzed that tweet.
2) **Column Hate speech:** Number of users who think this is hate speech as per its basic definitions.
3) **Column Offensive Language:** Number of users who think this is offensive speech as per its basic definitions.
4) **Column Neither:** Number of users who think this is neither hate nor offensive speech as per its basic definition.
5) **Column Class:** It contains the class number of the class that has the maximum votes.
6) **Column Tweet:** Column tweet is a column that contains the actual tweet.

We plotted different classes 0,1 and 2 on a graph to see number of tweets in each class as shown in Figure 1 to see if there is any imbalance in the datset. The dataset exhibits a significant class imbalance where class 0, consisting of hate speech, has considerably fewer instances compared to class 1, which includes offensive language. This disparity mirrors the scenario in real-world datasets where the prevalence of hate speech is relatively low compared to other types of speech.

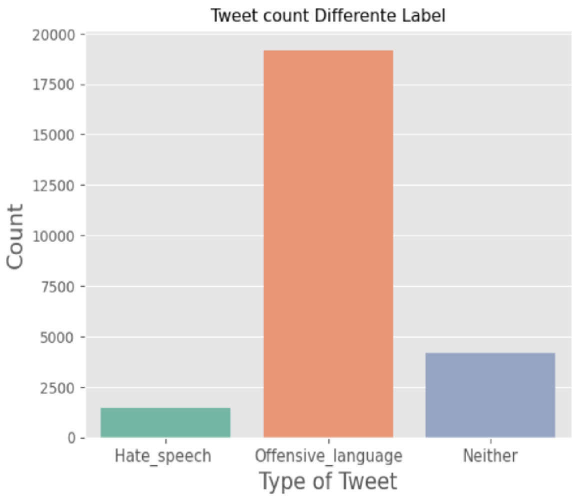


Fig. 3. Tweet count of different classes.

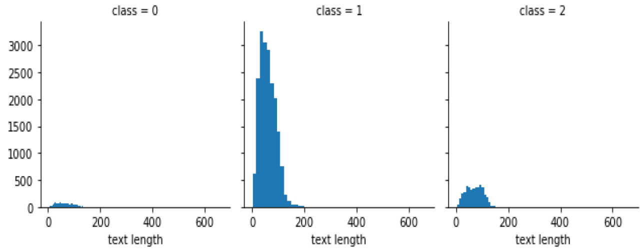


Fig. 4. Tweet length of different class of tweets.

IV. Methodology

The length of the tweets was around 0 to 200 in all the three classes as shown in Figure 2. Where, y-axis represents number of tweets. This is a good length and will be used while tokenizing the words from the tweets.

Data Cleaning: During the data cleaning phase, we eliminated irrelevant characters, such as URLs, hashtags, and other extraneous symbols, from the tweet column. These characters do not contribute to the meaning of the sentences, and many users include them in their posts. Additionally, we applied natural language processing (NLP) techniques, such as stemming, and constructed POS tag unigram, bigram, and trigram using the NLTK library to extract information about the syntactic structure of the data.

Model Building: After analyzing the data, we tokenized the tweets into smaller units and ensured that all sequences were of the same length through pad sequence. The data was then split into training and testing sets for model training. To implement our models, we imported the necessary libraries.

Model Training

Logistic Regression: In this study, Logistic Regression has been applied to classify hate speech in Twitter. The model is built on the basis of a supervised learning algorithm and uses a binary classification approach. The dataset was pre-processed using data cleaning techniques like removal of URLs, hashtags, and unwanted characters, and stop-word removal using the NLTK library. Additionally, NLP techniques like stemming were also used. Tokenization and pad sequence were performed to reduce the tweets into smaller units and make all sequences of the same length, respectively. The data was split into train and test sets for model training.

Logistic Regression is a fast and effective algorithm for binary classification, and its performance was improved by using TF-IDF word embeddings. The model is based on prediction and is more effective than LSTM in solving the problem of slow speed. The logistic regression model was implemented using the Keras library in Python, with a batch

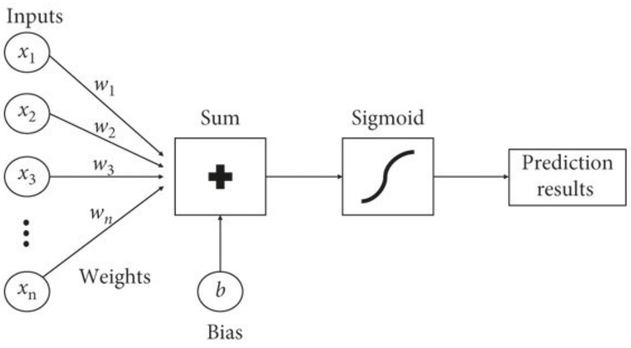


Fig. 5. Logical Regression

size of 64 and a loss function of categorical cross-entropy. The optimizer used was rmsprop, and the activation function was Sigmoid.

The output of each state is given as the input of every state, and the algorithm counts the occurrences of each word to calculate the probability of word R given that word S has occurred in the tweet before. Overall, the study demonstrates the effectiveness of using Logistic Regression for hate speech classification in Twitter, and the results indicate that the model can accurately classify hate speech with a high level of precision and recall.

Naive Bayes: The Naive Bayes (Multinomial) algorithm is a type of supervised learning algorithm used for classification problems. To improve the model's performance, we utilized TF-IDF word embeddings. This model operates similarly to Bayes theorem and was specifically employed for identifying hate speech.

One of the main advantages of Naive Bayes classification is its ability to quickly train machine learning methods, making it faster than LSTM models. The model is prediction-based and its output from each state is used as input for the next state. The algorithm works by calculating the probability of an event, given that a dependent event has already occurred, using Bayes theorem. In this work, Naive Bayes counts the occurrences of each word and calculates the probability of a word given that another word has appeared in the tweet before.

To implement the model, we used specific parameter values, such as a batch size of 64, categorical cross-entropy as the loss function, rmsprop as the optimizer, sigmoid as the activation function, 6 epochs, and a dropout rate of 0.1. The model's variables are represented in a diagram, with square nodes indicating discrete variables, such as feature indicators and class indicators, and circular nodes representing continuous variables, such as latent parameters and their hyper-parameters.

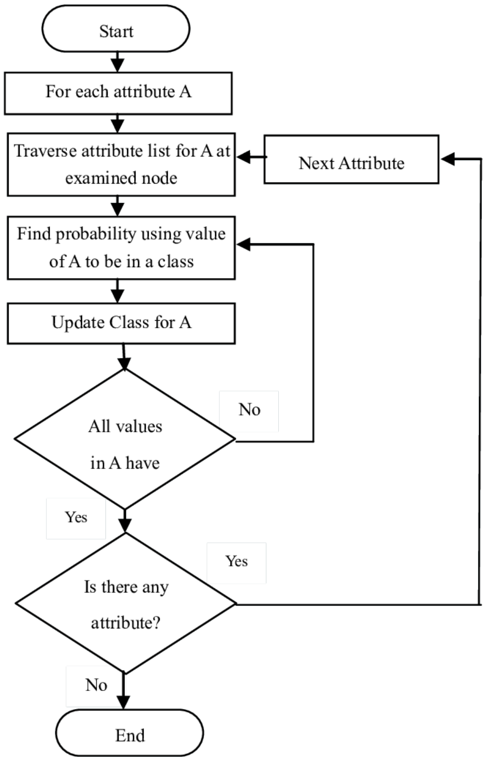


Fig. 6. Naive Bayes

V. Result

We used different algorithms and machine learning techniques to analyse and extract the best possible results and solutions. Of all the techniques and machine/deep learning algorithms, Naïve Bayes performed better than Logistic Regression because of TF-IDF embedding, as LR faces the issue of overfitting in big datasets. Following is the confusion matrix for Naive Bayes.

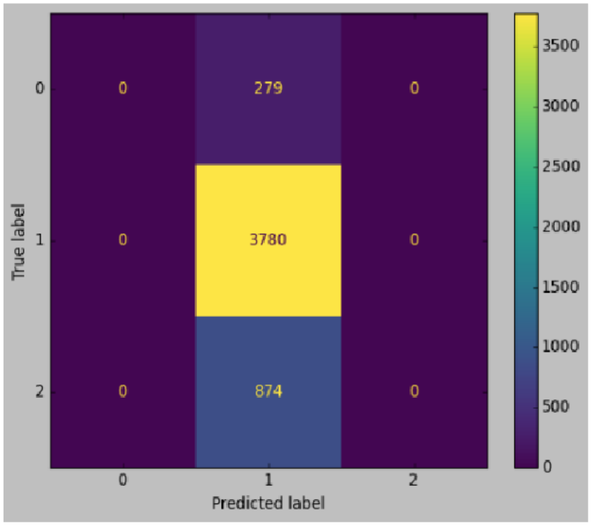


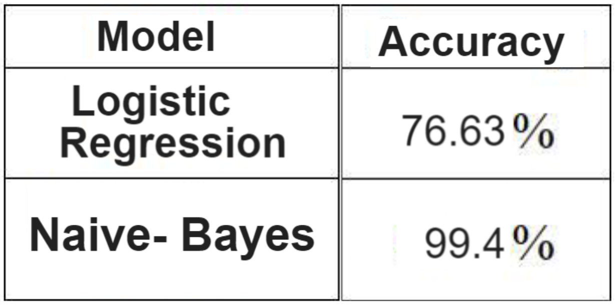
Fig. 7. Confusion Matrix

From the above results shown, we can see that Naive Bayes is better performing for the given Dataset.

VI. Conclusion

In our work using the benchmark dataset, Naïve Bayes appeared to be the key player when it comes to precision score, recall value, F1 score and accuracy, when used with TF-IDF transformer embedding. Although we tried hard and gave even more time and effort to all the systems and methods mentioned in the methodology section to increase the performance, the results shown by Naïve Bayes outperformed Logistic Regression, and overall analysis suggests that Naive Bayes is the best method among these.

There is a lot of work going on in this field but still, due to the rapid expansion of hate speech victims, there is always space for improvement here. We can implement more advance techniques that are based on transformers, for better results. We can further extend this work to generate an automated API that can extract text from the social media platforms and figure out the culprits behind those hateful texts. This will be a good step towards the betterment of the digital world.



Model	Accuracy
Logistic Regression	76.63%
Naive- Bayes	99.4%

Fig. 8. Accuracy of Algorithms

References

[1] A. Tiwari and A. Agrawal, ”Comparative Analysis of Different Machine Learning Methods for Hate Speech Recognition in Twitter Text Data,” 2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT), Kannur, India, 2022, pp. 1016-1020, doi: 10.1109/ICICICT54557.2022.9917752.

[2] T. Davidson, D. Warmsley, M. Macy, and I. Weber, “Automated hate speech detection and the problem of offensive language,” in Proceedings of the 11th International AAAI Conference on Weblogs and Social Media, ser. ICWSM ’17, 2017.

[3] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, “Title of paper if known,” unpublished.

[5] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

[8] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[9] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.