International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

HATE SPEECH DETECTION SYSTEM USING DEEP LEARNING

Mr. Rahul Wankhede , Sunil chavhan** Sujit jadhav, Mr Shubhum pawale, Ms. Nikita kolambe

[1234]B.E. Students, Department of Computer Engineering, Marathwada Mitra Mandal College of Engineering, Pune
[5]Assistant Professor, Department of Computer Engineering, Marathwada Mitra Madal College of Engineering,Pune

ABSTRACT

The growing use of social media and information sharing has provided significant benefits to humanity. However, this has resulted in a number of issues, including the spread and sharing of hate speech messages. To address this emerging issue on social media sites, recent studies used a variety of feature engineering techniques and machine learning algorithms to detect hate speech messages on various datasets. To the best of our knowledge, no study has been conducted to compare the various feature engineering techniques and machine learning algorithms in order to determine which feature engineering technique and machine learning algorithm outperform on a standard publicly available dataset. As a result, the purpose of this paper is to contrast the performance of three feature engineering techniques .Natural language Processing ,machine learning

***Keyword:** NLP, Machine learning.*

1. INTRODUCTION

Hate speech has increased in both in-person and online communication in recent years. Social media and other online platforms play a significant role in the breeding and spread of hateful content, which eventually leads to hate crime. According to recent surveys, the rise in online hate speech content has resulted in hate crimes such as Trump's election in the United States, the Manchester and London attacks in the United Kingdom, and the terror attacks in New Zealand. The European Union Commission has taken a number of steps, including legislation, to address the negative consequences of hate speech. Recently, the European Union Commission compelled social media platforms to sign an EU hate speech code that requires hate speech content to be removed within 24 hours. However, the manual process of identifying and removing hate speech content is time consuming and labour intensive. Because of these concerns and the prevalence of hate speech content on the internet, there is a strong case for automatic hate speech detection. In summary, we discuss the challenges and approaches to automatic hate speech detection, such as competing definitions, dataset availability and construction, and existing approaches. We also propose a new approach that outperforms the state of the art in some cases and discuss remaining shortcomings..

2. LITERATURE SURVEY

[1] Separating hate speech from other forms of abusive language is a major difficulty for automatic hate speech detection on social media. Because lexical detection approaches label all communications containing specified terms as hate speech, past work employing supervised learning has failed to distinguish between the two groups, they have low precision. We gathered tweets containing hate speech terms using a crowd-sourced hate speech lexicon. We used crowd sourcing to categorize a sample of these tweets into three categories: hate speech, offensive language only, and neither. To differentiate between these many groups, we train a multi-class classifier. An examination of the predictions and errors reveals when we can reliably distinguish hate speech from other offensive language and when it is more difficult to do so. We discovered that racist and homophobic tweets are more likely to be labelled as hate speech, while sexist tweets are more likely to be labelled as offensive. It's also more difficult to categorise tweets without obvious hate phrases.

[2] Understanding emotions is a fundamental part of personal development and evolution, and it's also a crucial component of human intelligence emulation. Emotion processing is vital not only for AI growth, but also for the closely linked problem of polarity identification. The ability to automatically capture public sentiments about social events, political movements, marketing campaigns, and product preferences has piqued interest in both the scientific community and the business world, thanks to the exciting open challenges and the remarkable marketing and financial market prediction implications. As a result, the fields of emotional computing and sentiment analysis have emerged, which use human-computer interaction, information retrieval, and multimodal signal processing to extract people's attitudes from the ever-growing amount of data available online social data.

[3] Hate speech identification on Twitter is essential for applications such as extracting controversial events, constructing AI chatterbots, content recommendation, and sentiment analysis. This assignment is defined as the ability to classify a tweet as racist, sexist, or neither. This undertaking is difficult due of the intricacy of natural language constructs. To manage this complexity, we conduct extensive research with

several deep learning architectures to learn semantic word embeddings. On a 16K annotated tweets benchmark dataset, we found that deep learning methods outperform state-of-the-art char/word n-gram algorithms by 18 F1 points. complexity.

[4] Hate speech is defined as any communication that disparages a target group of people on the basis of a trait such as race, colour, ethnicity, gender, sexual orientation, nationality, religion, or another feature. Hate speech is constantly increasing as a result of the tremendous surge in user-generated web material on social media. Interest in online hate speech identification, and especially the automation of this activity, has expanded in recent years, as has the societal impact of the phenomena. This study describes a hate speech dataset made up of thousands of sentences that were manually classified as containing or not containing hate speech. The sentences were taken from Stormfront, a white supremacist discussion board. To complete the hand labelling, a bespoke annotation tool was created. assignment that allows annotators to choose whether or not to read the context of a sentence before labelling it, among other things. The publication also includes a thorough qualitative and quantitative analysis of the dataset as well as multiple baseline tests with various categorization models. The data is open to the public.

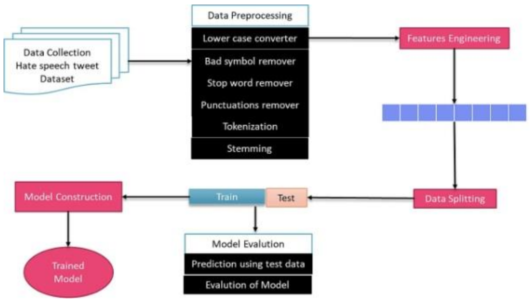
3. PROPOSED SYSTEM

1) Hate speech will be automatically identified, allowing the platform to detect and delete hate speech much more rapidly and efficiently.

2) Among the different machine learning methods, deep learning, a subset of machine learning, is widely used in Natural Language Processing (NLP) to address the problem of text classification.

A) SYSTEM ARCHITECTURE

This proposed system is used to detect hate speech with help of natural language processing and machine learning algorithms .For this system we required one dataset which contains large number of tweets. In this dataset 25000 tweets are there. Then Some data preprocessing is needed because machine needs clean data to give better output .after data preprocessing feature extraction is needed to build machine learning model.



An Algorithm Used

1. Decision tree

Decision tree Decision Tree is a supervised learning technique that can be applied to classification and regression problems; however it is most commonly employed to solve classification problems. Internal nodes represent dataset attributes, branches represent decision rules, and each leaf node provides the conclusion in this tree-structured classifier. The Decision Node and the Leaf Node are the two nodes of a Decision tree.

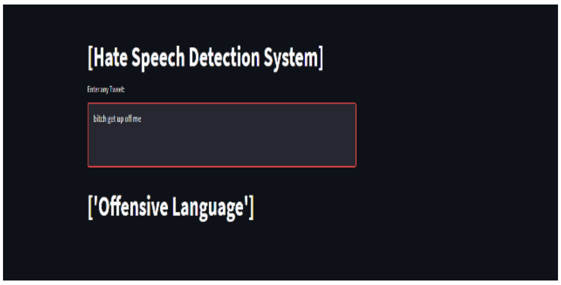
Figure 1: System Architecture

4. RESULT

The below snapshot define the working of the system. The performance stated below.

1) The text type by the user in the message box the text is categorize into hate speech or no hate .

2) The machine learning model given the accuracy of 87.



[Hate Speech Detection System]

Enter any Tweet:

bitch get up off me

['Offensive Language']

5. CONCLUSION

Finally, we discovered some chances in the field during our research. We discovered a dearth of open source platforms that automatically classify hate speech, no comparison studies that synthesize the methodologies used thus far, and the majority of the study was conducted only in English as a result of our thorough literature review. We believe that the task's intricacy and subjectivity make the learning process less immediate, and that annotators will need more training.

FUTURE WORK

As a result, in communities working to limit hate speech, it's critical to ensure that definitions are explicit, with rules, examples, and more uniform procedures. Finally, the outcomes of the experiment can be improved by extending it in the future. Increase the number of instances from classes with low representation as a first priority. In addition, with our effort, we attempted to identify and address some of the major issues in the field of automatic hate speech in text. Without a doubt, this is a field that still need further research and presents unending obstacles.

REFERENCES

[1] T. Davidson, D. Bhattacharya, and I. Weber, ``Racial bias in hate speech and abusive language detection datasets,'' in Proc. 3rd Workshop Abusive Lang. Online, 2019, pp. 2535

[2] O. de Gibert, N. Perez, A. García-Pablos, and M. Cuadros, ``Hate speech dataset from a white supremacy forum,'' in Proc. 2nd Workshop Abusive Lang. Online (ALW2), 2018, pp. 1120.