HATE SPEECH DETECTION BY MACHINE LEARNING

Anushka Thapliyal, Ayush Singh, Ritesh Kumar Singh
Student, ABES Engineering College, Ghaziabad, India
Student, ABES Engineering College, Ghaziabad, India
Student, ABES Engineering College, Ghaziabad, India
anushka.20b1541016@abes.ac.in, ayush.20b1541041@abes.ac.in, ritesh.20b1541002@abes.ac.in

Abstract – In today's world, as online content continues to grow, so does the spread of hate speech. Online toxic discourses could produce conflicts among various groups. Hate speech is complex and multifaceted harmful or offensive content targeting individuals or groups. We identify and examine problems came in our path while detection in text. Among these difficulties are subtleties in. We propose a multi-view SVM approach that achieves near state-of-the-art interpretable decisions than neural methods. We also discuss both technical and practical challenges that remain for this task.

In recent years, the increasing propagation of hate speech on social media and the urgent need for effective countermeasures have drawn significant investment from governments, companies, and researchers. Nowdays a very large data is produced every second of every day on social platform which consists of both good and bad speech. The main motive is to find the hate speech that is used in social media platforms. A large number of methods have been developed for automated hate speech detection online. This aims to classify textual content into non-hate or hate speech, in which case the (i.e., non-hate v.s. hate). In this work, we argue for a focus on the latter problem for practical reasons. that is difficult to discover. We then propose Deep Neural Network structures serving as feature extractors that are particularly effective for capturing the semantics of hate speech. Our methods F1, or 8 percentage points in the more challenging case of identifying hateful content.

Keywords: : hate speech, classification, neural network, CNN, GRU, skipped CNN, deep learning, natural language processing

INTRODUCTION

Hate crimes are unfortunately nothing new in society. Hate speech is a poisonous discourse that can swiftly spread on social media or due to prejudices or disputes
between different groups within and across countries. A hate crime refers to crimes committed against a person due to their actual or perceived affiliation with a specific group. For instance, suspects in several recent hate-related terror attacks had an extensive social media

history of hate related posts, suggesting that social media contributes to their radicalization [1, 2]. The protected characteristics of Facebook define hate speech as an attack on an individual's dignity, including their race, origin, or ethnicity. According to Twitter policies, tweets should not be used to threaten or harass others due to their ethnicity, gender, religion, or any other factor. In addition to age, caste, and handicap, YouTube also censors content that promotes violence or hatred toward certain persons or groups. In some cases, social media can play an even more direct role; video footage from the suspect of the 2019 terror attack in Christchurch, New Zealand, was broadcast live on Facebook. Vast online communication forums, including social media, enable users to express themselves freely, at times, anonymously. While the ability to freely express oneself is a human right that should be cherished, inducing and spreading hate towards another group is an abuse of this liberty.

For instance, The American Bar Association. Due to the societal concern and how widespread hate speech is becoming on the Internet [7], there is strong motivation to study automatic detection of hate speech. By automating its detection, the spread of hateful content can be reduced. Detecting hate speech is a challenging task, however. First, there are disagreements in how hate speech should be defined. This means that some content can be considered hate speech some and not to others, based on their respective definitions. We start by covering competing definitions, focusing on the different aspects that contribute to hate speech. We are by no means, nor can we be, comprehensive as new definitions appear regularly. Our aim is simply to illustrate variances highlighting difficulties that arise from such.

DEFINING HATE SPEECH

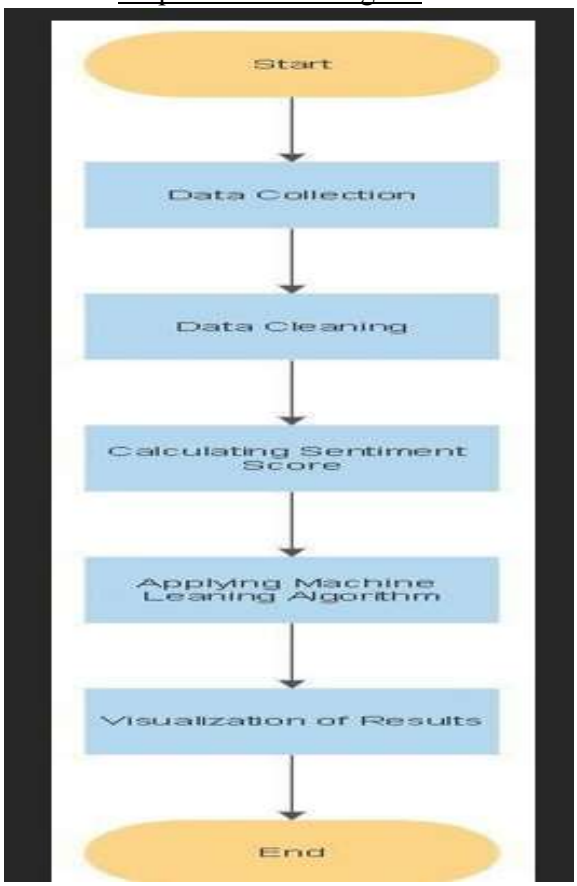
Hate crime Similarly, we opt not to propose a specific definition, but instead examine existing definitions to gain insights into what typically constitutes hate speech and what technical challenges the definitions might bring. We summarize leading definitions of hate speech from varying sources, as well as some aspects of the definitions that make the detection of hate speech difficult.

1. Encyclopedia of the American Constitution: "Hate speech is speech that attacks a person or group on the basis of attributes such as race, religion, ethnic origin, national origin, sex, disability, sexual orientation, or gender identity." [13] 2. and serious disease or disability. We also provide some protections for immigration status. We define attack as violent or dehumanizing speech, statements of inferiority, or calls for exclusion or segregation." [4] 3. Twitter: "Hateful conduct: You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease." [6] 4. Davidson et al.: "Language that is used to expresses hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group." [9] 5. de Gilbert et al.: "Hate speech is a deliberate attack directed towards a specific group of people motivated by aspects of the group's identity." [14] 6. Fortuna et al. and Facebook definitions. Fortuna et al.'s definition specifically calls out variations in language style and subtleties. Hate speech is abusive or threatening or writing that expresses prejudice on the basis of ethnicity , religion , sexual orientation . whenever a government passes a bill, it is important for the government to know that if the bill passed is being liked by all or not. In 2021, The Modi Government passed the farmer's bill, where the reaction from the public was mixed. Some people were happy about it, and some were not. There were people who were using hate speech and offensive words. This can be challenging, and goes beyond what conventional text based classification approaches are able to capture. Fortuna et al.'s definition is based on an analysis of the following characteristics from other definitions [8]: A particular problem not covered by many definitions relate to factual statements. For example, "Jews are swine" is clearly hate speech by most definitions (it is a statement of inferiority), but "Many Jews are lawyers" is not.

METHODOLOGY

Methodology is a process of that shows how an algorithm works and consist of various flowcharts representing the way our work is going to happen in various stages.

Proposed Model Diagram



The proposed methodology related to our project is given below:
Step 1: Identify the abusive words in the tweet. Tweets

are extracted from the dataset present on the Kaggle.

Step 2: The preprocessing of the dataset is done. It involves the following steps:

- Removal of capitalization
- Remove whitespace with a single space
- Removal of punctuations and numbers
- Removal of extra spaces
- Removal of links[https://abc.com]

Step 3: Analyzing the polarity of the dataset.

Step 4: Giving the step 3 output in machine learning algorithm and analyze it to find the algorithm with best accuracy.

Step 5: The result is obtained.

DATASETS

Datasets have targeted numerous hate speech categories. However, examples could be unclear in several datasets, such as Waseem's dataset [27] or hierarchical datasets including the dataset of Basile et al. [58 In addition, approximately 60% of dataset creators found inter-annotator agreement [1]. Therefore, a useful predictive hate speech detection model requires relevant and non-obsolete datasets. The maturity of datasets is regarded as a one-of-a-kind task for superior quality systems. According to Koco et al. [129], separating annotator groups has a more significant impact on the performance of hate detection systems. They also stated that group consensus impacts the quality of recognition .hate speech is difficult to measure [130]. Our analysis showed that research requires more robust, trustworthy, and large datasets due to the wide applications of hate speech detection. The dataset of Shibly et al. [65] developed a robust and colossal dataset.

Those datasets backed by conference or workshop series, such as Hate Speech and Offensive Content Identification in Indo-European Languages (HASOC) [19,38] and Sem Eval, are probably among the most popular datasets. HASOC is divided into three subtasks: the first focuses on identifying hate speech and offensive language (sub-task A); the second focuses on identifying the type of hate speech (sub-task B); and the third focuses on identifying the hate speech's target group (or persons). The Sem Eval Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter continues to focus on numerous academics in the Sem .Eval series [83,113]. On the other hand, researchers have paid close attention to Sem Eval 2019 Task 6 [18]—Offense Eval: Identifying and Categorizing Offensive Language on Social Media. The Offensive Language Identification Dataset (OLID) [101], which contains over 14,000 English tweets, is the most recent.

• HatebaseTwitter [9]. One Twitter dataset is a set of 24,802 tweets provided by Davidson, et al [9]. Their procedure for creating the dataset was as follows. First they took a hate speech lexicon from Hatebase [16] and searched for tweets containing these terms, resulting in a set of tweets from about 33,000 users. Next they took a timeline from all these users resulting in a set of roughly 85 million Tweets. From the set of about 85 million tweets, they took a random sample, of 25k tweets, that contained terms from the lexicon

• WaseemA [17]. Waseem and Hovy also provide a dataset from Twitter, consisting of 16,914 tweets labeled

as racist, sexist, or neither [17]. They first created a corpus of about 136,000 tweets that contain slurs and terms related to religious, sexual, gender, and ethnic minorities.

• GermanTwitter [11]. As part of their study of annotator reliability, Ross, et al. created a Twitter dataset in German for the European refugee crisis [11]. It consists of 541 tweets in German, labeled as expressing hate or not.

ALGORITHMS USED

In this program, we have used Decision Tree Classifier. It is a supervised learning algorithm that is used to produce a result which is either "Yes" or "No". there is a decision node in the algorithm which is represented by a square box. There are leaf nodes which are shown by circle . Decision Tree is a tree like structure which used supervised learning algorithm. It is used in both classification and regression, but it is mainly used in classification purposes.

In Supervised learning , the approach is domain dependent since it relies on a manual labeling of a large volume of text. Labeling task is time and effort consuming but it is more efficient for domain-dependent events. Most of the approaches used for hate speech detection tasks are supervised methods. For instance, Burnap and Williams [30] have used several supervised classifiers to detect hate speech

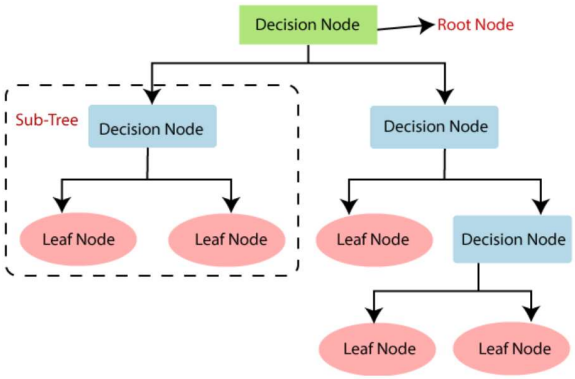
in twitter, their results showed that all classifiers have performed the same but .

Decision Tree Classifier

In this program, we have used Decision Tree Classifier. It is a supervised learning algorithm that is used to produce a result which is either "Yes" or "No". there is a decision node in the algorithm which is represented by a square box. There are leaf nodes which are shown by circle . Decision Tree is a tree like structure which used supervised learning algorithm. It is used in both classification and regression, but it is mainly used in classification purposes.

value of the target variable, or when splitting no longer adds value to the predictions. parameter setting, and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle high-dimensional data. In general decision tree classifier has good accuracy. Decision tree induction is a typical inductive approach to learn knowledge on classification.

Supervised learning is also known as supervised machine learning which is sub category of the machine learning algorithms and artificial intelligence . It uses labelled data sets to train algorithm that's is used to classify the data all predict the outcomes accurately.



RESULT

We can infer that our model for **Hate speech detection** performs with an accuracy of 87 percent.

The result we got from analyzing the tweet is given below in Fig.

Fig. shows that tweet is an hate speech as it contain harsh words.

CONCLUSION

Thus, after performing decision tree classifier, which is used in supervised learning algorithm, we get to know that with an accuracy of 87%, our result is “hate speech”. Similarly, we can use this algorithm for all the datasets that are taken from all the spocial media platforms like twitter, facebook and Instagram

REFERENCES

1. Hate Speech—ABA Legal Fact Check—American Bar Association;. Available from: https:// abalegalfactcheck.com/articles/hate-speech.html.

2. Guardian. Anti-muslim hate crime surges after manchester and london bridge attacks, Last accessed: July 2017, https://www.theguardian.com.

3. Hateful conduct policy;. Available from: https://help.twitter.com/en/rules-and-policies/hateful-conductpolicy.

4. Dong YS, Han KS. Boosting SVM classifiers by ensemble. In: Special interest tracks and posters of the 14th international conference on World Wide Web. ACM; 2005. p. 1072–1073.

5. Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of Tricks for Efficient Text Classification.

6.Davidson T, Warmsley D, Macy MW, Weber I. Automated Hate Speech Detection and the Problem of Offensive Language. ICWSM. 2017;.

7. Vig J. Visualizing Attention in Transformer-Based Language Representation Models. arXiv preprint arXiv:190402679. 2019;.

8. Billy Chiu, Anna Korhonen, and Sampo Pyysalo. Intrinsic evaluation of word vectors fails to predict extrinsic performance. In Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP, pages 1–6. Association for Computational Linguistics, 2016. doi:10.18653/v1/W16- 2501.

9. Hagen M, Potthast M, Bu¨chner M, Stein B. Webis: An Ensemble for Twitter Sentiment Detection. In: SemEval@NAACL-HLT; 2015.

10. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed Representations of Words and Phrases and their Compositionality. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, editors. Advances in Neural Information Processing Systems 26. Curran Associates, Inc.; 2013. p. 3111–3119.

11. Yashar Mehdad and Joel Tetreault. Do characters abuse morethan words? In Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 299–303. Association for Computational Linguistics, 2016. doi:10.18653/v1/W16-3638.

12. David Robinson, Ziqi Zhang, and John Tepper. Hate speech detection on Twitter: feature engineering v.s. feature selection. In Proceedings of the 15th Extended Semantic Web Conference, Poster Volume, ESWC’18, pages 46–49, Berlin, Germany, 2018. Springer. doi:10.1007/978-3-319-98192-5_9.

13. Zhao J, Xie X, Xu X, Sun S. Multi-view learning overview: Recent progress and new challenges. Information Fusion. 2017;.