Hate Speech Detection on Social Media Using Machine Learning Algorithms

Rupesh Chaudhari[1,*], Ritik Gade[1], Pranav Gawali[1], Mangesh Gite[1], A. B. Pawar[1]

[1]Computer Engineering, Sanjivani College of Engineering, Kopargoan, Savitribai Phule Pune University, India
Emails: rupeshchaudhari2151@gmail.com, ritikgade22@gmail.com, pranavgawali2510@gmail.com, mangeshgite9@gmail.com, pawaranilcomp@sanjivani.org.in

Abstract

There is an enormous growth of social media which fully promotes freedom of expression through its anonymity feature. Freedom of expression is a human right but hate speech towards a person or group based on race, caste, religion, ethnic or national origin, sex, disability, gender identity, etc. is an abuse of this sovereignty. It seriously promotes violence or hate crimes and creates an imbalance in society by damaging peace, credibility, and human rights, etc. To overcome this problem, the hate speech detection model is made which will classify the speech and if the speech used by user is containing hate word, it will be detected and system will sent an alert message to user about it. In order to solve various hate speech problems we use some of the machine learning algorithms such as logistic regression and random forest. If user disrupts cyber guidelines, then strict action shall be taken and user's account will be ban forever. This help to reduce cyber crimes in effective and efficient manner.

Keywords: Machine learning; Hate speech; Natural language processing; Data pre-processing; Random forest; Logistic regression; Hate word classification

1. Introduction

There is an enormous growth of social media which fully promotes freedom of expression through its anonymity feature. Because of that detecting hate speech in social media discourse is quite essential but a complex task. From this project, we can addressing different categories of hate separately, accurately predict their different forms, by exploring a group of text mining features. Building a system that can detect the hate spreader words/speech from social media platforms. Those hate speeches are further classified in different form of hate speeches like gender, caste, nation, etc. using ML algorithms. Main objective of this prediction model is to detect the hate spreader speech from social media platforms. We predict what has been going on social media and take precautions regarding hate speeches. We can stop spreading hate about anything, person or other stuff so the society will live peacefully.

2. Literature survey

For this project, we referred some papers published in recent years and some websites to get idea about what and how we can proceed in our project. Some of them are given as follows:

Doi : https://doi.org/10.54216/JCHCI.020203
Received:December28, 2021 Accepted:April2, 2022

"A Survey on Hate Speech Detection using Natural Language Processing" paper published in 'Proceedings of the fifth international workshop on natural language processing for social media' in 2017. 'Anna Schmidt' and 'Michael Wiegand' are the authors of this paper. In this paper, they presented a survey on the automatic detection of hate speech. They framed it as a supervised learning problem. They achieve it by fairly generic features, such as bag of words or embeddings, systematically classification performance [1].

"Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection" is published in IEEE in 2018. This paper proposed by 'Hajime Watanabe', 'Mondher Bouazizi' and 'TomoakiOhtsuki'. While most of the online social networks and micro blogging websites forbid the use of hate speech, but the size of these networks and websites make it almost impossible to control all of their content. In this paper, they propose an approach to detect hate expressions on Twitter. The approach is based on unigrams and patterns that are automatically collected from the training set. These patterns and unigrams are later used, among others, as features to train a machine learning algorithm [2].

"The Ongoing Challenge to Define Free Speech" article is published in 'Human Rights Magazines' in 2018. 'Wermiel SJ' is author of this article. In this they stated that "The controversy over what many call "hate speech" is not new, nation experiences the Black Lives Matter movement because of hate speech which promotes national dialogue about racism, sexual harassment, and more which is quite dangerous and the individual practising the hate speech is being prohibited or punished in (USA) according to the law."[3].

In "Youtube" website they clearly stated strict policies against hate speeches in 2021. Hatred speech which Encourage violence against individuals or groups based on any of the attributes (sex, gender, racial, religion etc) such kind of speech is reported and block by YouTube. If we violate rules they may give us warning or our channel should be banned/ blocked forever [7].

"Un-Compromised Credibility: Social Media based Multi-Class Hate speech Classification for Text" was proposed by Khubaib Ahmed Qureshi , Muhammed Sabih in 2021. Detecting hate speech in social media is quite essential but because of increase in data it get difficult to detect . In this paper, they aims to accurately predict different forms of hate speeches, by exploring a group of text mining features. The highest categories of hate crimes reported by the FBI, are based on race, ethnicity, religion, and sexual orientation. Therefore all these categories are primarily selected for dataset's compilation. They use multiple algorithms for this like Logistic regression, Random forest, SVM, etc.[8].

"Analyzing the hate and counter speech accounts on Twitter" was proposed by Binny Mathew, Navish Kumar, Ravina, Pawan Goyal, Animesh Mukherjee in 2018. In this paper, they analyze hate speech and the corresponding counters on Twitter. They perform many linguistic and psycholinguistic analysis on some user accounts on Twitter and obverse that counter speakers use or apply many different strategies depending upon which community peoples they wanted to target. They also find that the hate tweets by verified accounts get much more viral as compared to a tweet by a non-verified account. While the hate users mostly use words about envy, hate, negative emotion, swearing terms, ugliness, the counter users use more words related to government, law, political/society leaders. So they build a supervised model for classifying the hateful and counter speech accounts on Twitter[9].

Doi : https://doi.org/10.54216/JCHCI.020203
Received:December28, 2021 Accepted:April2, 2022

3. System Architecture

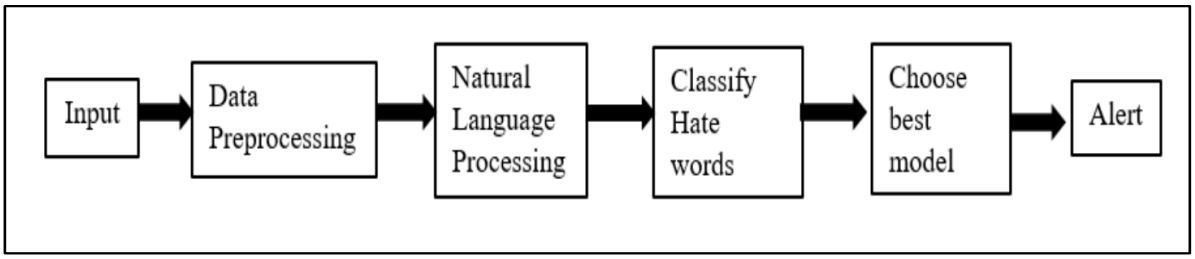

Figure 1: Proposed System Architecture

4. Methodology

- **Input:**The dataset is imported from "kaggle.com" or from developed web applications and will import those using pandas.
- **Data Preprocessing:** To transform raw data into understandable format using possible pre-processing techniques. Stop ward Removal Concept will be applied on the data which removes all punctuation and spaces to rearrange it into proper manner.
- **Natural Language Processing (NLP):** We will use TFIDF-Vectorizer concept of NLP for converting a text format data into numeric value. It will give each hate speech word a specific vector value. Formula by which TFIDF convert text value to numeric value :

$$TF(Term\ Frequency) = \frac{no\ of\ repeated\ value\ in\ speech}{no\ of\ words\ in\ speech}$$

$$IDF(Inverse\ Document\ Frequency) = log\ \frac{no\ of\ speech}{no\ of\ speech\ containing\ that\ word}$$

$$Numeric\ Value = TF * IDF$$

- **Classify Hate Speech:** Algorithms will be used to classify hate text using data, algorithms like logistic regression or random forest.
- **Choose the Best Model:** By using confusion matrix, we will compare the accuracy percentage of results of Logistic Regression and Random Forest. A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. After that we will choose the best model.
- **Alert:** After detecting hate speech system give alert to user that in the given input, if system found some hate content it will sent the warning alert to user.

5. Algorithm

1. **Random Forest:** Random Forest is algorithm which makes decision on data as 0 if hate detected and 1 for normal speech and then determined the result. The main advantage of using Random Forest is that it has high accuracy and less variance than a single decision tree. The Data Pre-processing is done and then based on the parameters in the dataset.
   Step 1 : In Random Forest n number of random records are taken from the data set having k number of records.
   Step 2 : Individual decision trees are constructed for every user's entered speech.

Doi : https://doi.org/10.54216/JCHCI.020203
Received:December28, 2021 Accepted:April2, 2022

Step 3 : Each decision tree will generate an output by predicting the hate speech category.
Step 4 : Final output is considered based on Majority Voting or Averaging for Classification and regression respectively.
Step 5 : At last, select the most voted prediction as the final category of hate speech predicted.

2. **Logistic regression:**Logistic regression model is a model for calculating probabilities between 0 and 1. We will label hate speech as 1 and normal speech as 0 and then determine the result.

6. Applications

This project can be used many ways. Some of them are listed below :

- To prevent war/conflict cause because of hate speech.
- To provide assistant to cyber cell in order to detect crime.
- To reduce racial, caste, religion, ethnic or national origin, sex related conflicts/Wars.
- To secure human rights.

7. Conclusion

Now-a-days, the use of Social Media is rapidly growing. In this situation, taking precautions about anyone can't get affected by social media is most important. So, this system will check text content posted on social platforms and will identify the hate words/speech from those posts. After that it will categorized using ML algorithms like Random Forest, Logistic Regression, etc. This will help to authorities of government or of social media companies to handle this kind of vulgar/hate content and will maintain dignity of individual one.

References

[1] Schmidt, Anna Wiegand, Michael, "A Survey on Hate Speech Detection using Natural Language Processing", 1-10. 10.18653/v1/W17-1101. URL : https://aclanthology.org/W17-1101 , (2017).
[2] H. Watanabe, M. Bouazizi 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, pp. 13825-13835, 2018, doi: 10.1109/ACCESS.2018.2806394.
[3] Stephen Wermiel, "The Ongoing Challenge to Define Free Speech", 43 Human Rights 82 (2018).
[4] MacAvaney S, Yao H-R, Yang E, Russell K, Goharian N, Frieder O, "Hate speech detection: Challenges and solutions", PLoS ONE 14(8): e0221152, https://doi.org/10.1371/journal.pone.0221152 , (2019).
[5] Vu, Xuan-Son Vu, Thanh Tran, Mai-Vu Le-Cong, Thanh Nguyen, Huyen, "HSD Shared Task in VLSP Campaign 2019:Hate Speech Detection for Social Good",(2020).
[6] 'Twitter' Website, 2021, The Twitter Rules, Retrieved from this site: https://support.twitter.com/articles/.
[7]'Youtube' Website, 2021, Hate speech, Retrieved from this site : https://support.google.com/youtube/answer/2801939?hl=en
[8] K. A. Qureshi and M. Sabih, "Un-Compromised Credibility: Social Media Based Multi-Class Hate Speech Classification for Text", in IEEE Access, vol. 9, pp. 109465-109477, 2021, doi: 10.1109/ACCESS.2021.3101977.
[9] Mathew, Binny, et al. "Analyzing the hate and counter speech accounts on twitter." arXiv preprint arXiv:1812.02712 ,url- https://doi.org/10.48550/arXiv.1812.02712 (2018).
[10] P. Kavitha , R. Subha Shini , R. Priya, "An Implementation Of Statistical Feature Algorithms For The Detection Of Brain Tumor", Journal of Cognitive Human-Computer Interaction, 2021, DOI: https://doi.org/10.54216/JCHCI.010202.

Doi : https://doi.org/10.54216/JCHCI.020203
Received:December28, 2021 Accepted:April2, 2022