International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 11 Issue: 11 | Nov 2024 www.irjet.net p-ISSN: 2395-0072

DETECTION OF POLITICAL HATE SPEECH FOR SOCIAL MEDIA MONITORING

Vaibhav Shah[1] and Soham Kulkarni[2] and Chinmay Deodhar[3] and Viraj Shah[4] and Priyanca Gonsalves[5]

[1] Dwakadas J. Sanghvi College of Engineering, Mumbai, India
[2] Dwakadas J. Sanghvi College of Engineering, Mumbai, India
[3] Dwakadas J. Sanghvi College of Engineering, Mumbai, India
[4] Dwakadas J. Sanghvi College of Engineering, Mumbai, India
[5] Dwakadas J. Sanghvi College of Engineering, Mumbai, India

---

Abstract - *Political hate speech is an expression that is conveyed from a person or a group towards other person or group as an attack using various media. It can be found in a variety of online and offline contexts, including social media, news articles, and public speeches. Political hate speech can result in various negative consequences like public unrest , riots and protests and increased criminal activity. This project will develop a novel approach to political hate speech detection that combines machine learning and NLP techniques. The approach mentioned in this work would identify and note the features that determine the hate speech. These features may include keywords, phrases, and sentiment. After the features are successfully extracted from the raw text , they are input to the machine learning models which will predict if the speech contains hate. The input data can be in any format like text, audio files, or even YouTube URLs and it will classify what type of emotion and the percentage of hate speech comes from it. The machine learning model mentioned in the work will be trained on a large publicly labelled dataset. The developed model will also predict the emotion of the given input sample and classify it to various parameters like joy , sadness, anger ,surprise , etc. Our proposed work will be based on a text dataset of political speeches and articles. The results of the work will be reviewed in terms of accuracy and flawlessness of the model to predict the correct label. The proposed approach will also be compared to existing hate speech detection systems. In this work we have implemented the task in different machine learning algorithms including adaboost , decision tree and random forest [1] [3] [5].*

Key Words: Political hate speech, Emotion Detection, Automatic detection, Machine learning, Natural language processing (NLP), Feature extraction, Sentiment analysis, Classification.

1.INTRODUCTION

The widespread nature of political hate speech presents a serious danger to individuals, communities, and democratic societies. This damaging type of expression can provoke violence, undermine social unity, and silencethose who are already marginalized. With the growth of online platforms and social media, the reach and influence of political hate speech have increased, making it more critical than ever to create effective strategies for detection and mitigation. The main goal of detecting political hate speech is to recognize and categorize instances of this harmful language across different forms of communication, such as social media posts, news articles, and public speeches [8].

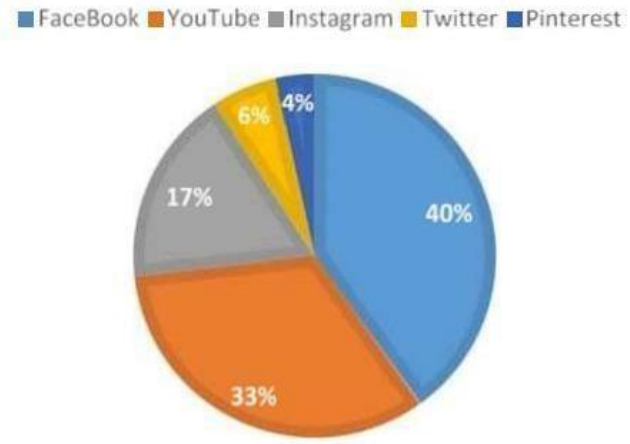

Fig. 1. Active user on social media

We can: Prevent and protect people or groups from harm, Prevent cyberbullying from occurring in the first place, help the police and courts: Hate speech can be traced in order to assist both the law enforcement and legal processin bringing a perpetrator of hate crime to justice [2]. In a nutshell, political hate speech reaches even the most sensible of individual's modes of detection, which in itself proves that individuals can be protected further. This has been interpreted to mean support for further investigation of the political hate speech through novel means that perpetuates further discussions of such continuations. [5].

2. LITERATURE REVIEW

2.1 Existing Work

Literature Related to Existing System : Abro et al. (2020) implemented Naïve Bayes model for hate speech detection and on a dataset labelled from Crowdflower and achieved 87% accuracy for the classification task [1].Its drawback is that it does not have any other algorithm except Naïve Bayes. Alaoui et al. (2022) implemented various algorithms including SVM, KNN, Decesion Tree, Adaboost and Naïve Bayes on a dataset containing 14500 tweets among which Naïve Bayes algorithm was highest performing with 79% accuracy [2]. Shubhang et al. (2023) achieved 77% accuracy on datasets collected from various sources like strimfront, kaggle, etc [3]. Md.G et al. (2021) implemented hate speech detection on a dataset obtained from twitter containing 24000 tweets and implemented many algorithms among which SVM was the highest performing [4].

Literature Related to Methodology/ Algorithms : Bothe et al. (2019) describes the methodology of conversion of audio or video to text or transcription of text from audio files and application of algorithms for the same [5]. Abro et al. (2020) has implemented hate speech classification in Naïve Bayes model with accuracy of 87% and has described the methodology for the preprocessing, tokenization and application of model for the dataset from twitter [1]. Ombui et al. (2019) have implemented the methodology for implementation of SVM, LSTM, Naïve Bayes with SVM having highest accuracy of 77%. Alaoui et al. (2022) have implemented the complete methodology for web scraping from twitter and application of data mining to classify the tweets as hate or non-hate. Shukla et al. (2022) have implemented the methodology for hate speech detection on hindi language tweets and posts from other social media datasets which is available on site codalabs [7].

Literature Related to Technology/ Tools/ Frameworks : Abro et al. (2020) implemented Naïve Bayes model for hate speech detection and on a dataset labelled from Crowdflower and achieved 87% accuracy for the classification task [1].Its drawback is that it does not have any other algorithm except Naïve Bayes. Alaoui et al.. (2022) implemented various algorithms including SVM, KNN, Decesion Tree, Adaboost and Naïve Bayes on a dataset containing 14500 tweets among which Naïve Bayes algorithm was highest performing with 79% accuracy [2].

2.2 Observations On Existing Work

Abro et al. (2020) implemented Naïve Bayes model for hatespeech detection and on a dataset labelled from Crowdflower and achieved 87% accuracy for the classification task.Its drawback is that it does not have any other algorithm except Naïve Bayes. Alaoui et al. (2022) implemented various algorithms including SVM, KNN, Decesion Tree, Adaboost and Naïve Bayes on a dataset containing 14500 tweets among which Naïve Bayes algorithm was highest performing with 79% accuracy.Ombuiet al.(2019) have implemented the methodology for implementation of SVM, LSTM, Naïve Bayes with SVM having highest accuracy of 77% [6].Md et al. (2021) implemented hate speech detection on a dataset obtained from twitter containing 24000 tweets and implemented many algorithms among which SVM was the highest performing [4]. Shukla, S., Nagpal, S. and Sabharwal, S. (2022) presents the model that can detect hate speech in real time has 2 stage architecture applying BERT encoder, using CNN followed by sigmoid layer. Dataset contains posts extracted from YouTube and Twitter [7].

3. METHODOLOGY

Methodology

A schematic representation of the system model for the proposed multi-modal hate speech detection is illustrated in Fig. 1. In the first stage of video data capturing, we process the data for several operations including image resizing, noise removal, and transformation of the data into image, text, and audio. The features of audio data are extracted from both the time and frequency domains before being forwarded the the model. Subsequently, images, text and audio of the extracted features will be supplied to the model. The model predicts whether the input is a hate speech or not and assigns the respective label to the input data. Subsequently, to establish the final decision prediction, a majority voting ensemble will be used in which the core decision basically is, two or more exposed modes of communication must be hate, in order for the final prediction to be so [8] [9].

Data Description and Pre-processing

We have assembled the necessary labeled data to put into the system for training which has two classes hatred and non-hatred. Hatred feelings and expressions relates to the feelings of anger, fear, aggression, disgust, dislike, encountered in violent communications or what research can consider the injurious or violent feelings at their most extreme level [7]. Conversely, non-hatred feelings or positive emotions such as those described by [8] tend to include the aspect of enjoyment or anything positive that one would want to indulge in such as among others pleasure, gladness, excitement etc. In total, we have 1051 video data collected from 3 movies, 2 web series, 3 hate speech, and 4 positive emotions in order to study hate and positive emotions addressing 80% for training and 20% for testing purpose.

```
In [4]: df['Emotion'].value_counts()

Out[4]:
joy        11045
sadness     6722
fear        5410
anger       4297
surprise    4062
neutral     2254
disgust      856
shame        146
Name: Emotion, dtype: int64
```

Fig 2 Emotion detection from text

Label	Percentage in data
Hate	30
Non - Hate	60
Offensive	10

Fig 3 . The distribution of hate speech and non hate speech

Feature extraction and feature selection:

Counter Vectorizer:

The bag-of-words (BOW) is the representation of text data that is converted to a vector form . This vector represents the input data according to how many times the word appears in the text .

Term Frequency-Inverse Document Frequency (TFIDF):

TFIDF is a mathematical representation of natural text in which the importance of each word is assessed according to the number of times the term appears in the text. The formula that is used to compute the tf-idf is:

$$TF - IDF = TF * IDF \quad (1)$$

4. SYSTEM ARCHITECTURE

The proposed system architecture for political hate speech detection from social media leverages a combination of web scraping techniques and advanced natural language processing (NLP) algorithms.

Data Collection Module

The data collection module obtains information from social media pages, news websites, articles, and other geotexts. An example of the techniques used involves web scraping for posts, comments, and articles relating to political discourse. Beautiful Soup, or Scrapy are other web scraping tools used to collect data from the specified websites in a manner that is within the required ethical standards and does not breach any privacy laws. The information collected is first processed and prepared before it is used in the next modules for data analysis [2].

Natural Language Processing (NLP) Module

The main task of the algorithms and the models included in the NLP module is the detection of hate-speech in text documents. The understanding of political discussion is supported by advanced NLP approaches, including sentiment analysis, entity recognition, and language model embeddings. Such language models, which can be used to embrace a wider range of political discourses, differentiate between the acceptable and unacceptable forms of expressing disagreement using politically charged language. The NLP module integrates context sensitivity, allowing the system to be customized for multiple political contexts and linguistic forms.

Machine Learning Models Used

Random Forest is also called an ensemble learning technique which builds numerous decision trees when being trained and can then use majority voting or averaging method when making predictions. The Decision Tree works under supervision by partitioning a set of data into disjointed subsets according to the input characteristics features and the output is in a tree-structured model of decisions and their outcomes. AdaBoost is a kind of ensemble learning and it is based on the concept of comb by many simple classifiers to build a strong one by re-weighting the incorrectly classified samples.

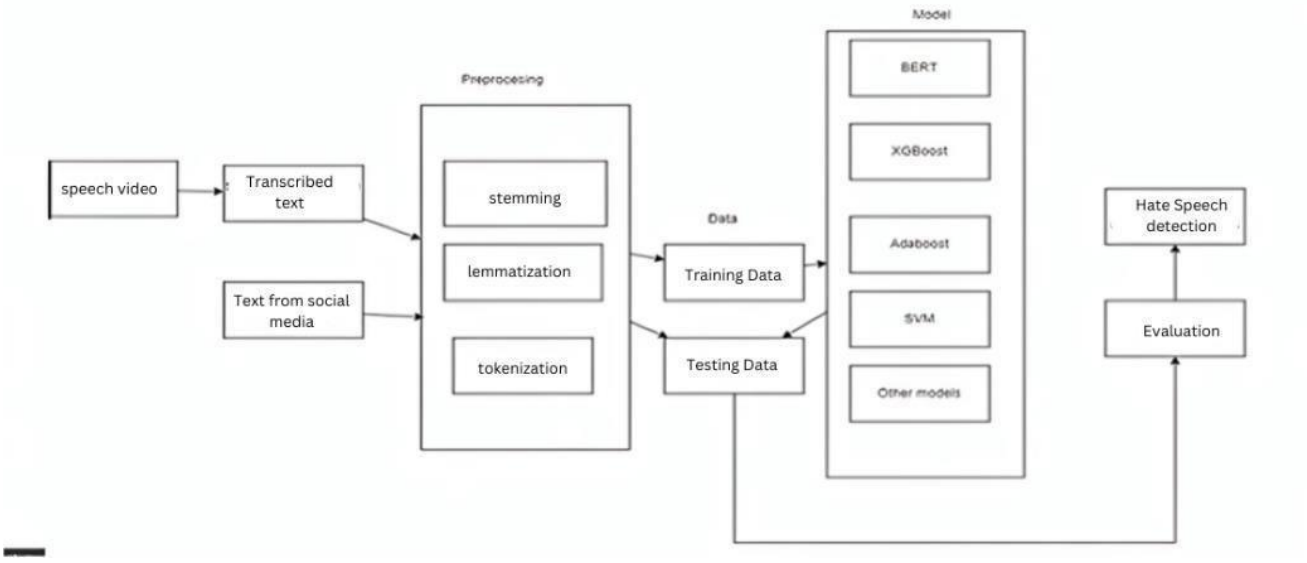

Fig. 4. System Architecture

Ethical Considerations

The proposed system architecture aims to detect political hate speech from social media platforms using web scraping techniques and advanced NLP algorithms. It incorporates ethical considerations to ensure fairness, transparency, and user privacy, and to mitigate potential biases and safeguard user rights. This comprehensive solution aims to create a safer online environment for political discourse.

5. RESULT

The study evaluated three machine learning models: Random Forest, Decision Tree, and AdaBoost, for a hate speech detection system. AdaBoost, which combined weak learners, was the most robust in accurately classifying hate speech instances, outperforming the other models. Their performance was analyzed by confusion matrices to determine classification accuracy and errors for the Random Forest, Decision Tree, and AdaBoost models. In Random Forest accuracy was measured up to 90 percent and Decision Tree was measured up to 89 percent mobile. AdaBoost proved with 93 % accuracy, the way of how the weak learners can be combined to enhance the predictive capability.

6. CONCLUSIONS

Regarding the findings of this research, it is established as follows: This research aims to develop and assess machine learning such as Random Forest, Decision Tree, and AdaBoost to determine the existence of political hate speech in social media. These models as shown in the results had fair accuracy with AdaBoost having the highest at 93% while Random Forests had 90% and Decision Trees at 89% all infer that technology does not suffice for hate speech problems. The models were able to recognize different patterns in toxic content, however, they had issues with understanding context, or intent and interpersonal communication. In addition, the ethical implications of such models, including unduly bias training data and concerns of censorship significantly emphasize the need for a very open model that will also be continuously refined. Finally, we argue that although AI has great potential to address the problem of political hate speech, AI needs to be delegated with human supervision and accountability for the satisfactory solution.

REFERENCES

1. Abro, S., et al.: Automatic hate speech detection using machine learning: A comparative study. International Journal of Advanced Computer Science and Applications 11(8). doi:10.14569/ijacsa.2020.0110861 (2020).

2. Alaoui, S.S., Farhaoui, Y., Aksasse, B.: Hate speech detection using text mining and machine learning. International Journal of Decision Support System Technology 14(1), 1–20. doi:10.4018/ijdsst.286680 (2022).

3. Boishakhi, F.T., Shill, P.C., Alam, Md.G.: Multimodal hate speechdetection using machine learning. In: 2021 IEEE International Conference on Big Data (Big Data) [Preprint]. doi:10.1109/bigdata52589.2021.9671955 (2021).

4. Bothe, H.H.: Audio to audio-video speech conversion with the help of Phonetic Knowledge Integration. In: 1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation [Preprint]. doi:10.1109/icsmc.1997.638237 (no date).

5. Ombui, E., Muchemi, L., Wagacha, P.: Hate speech detection in code-switched text messages. In: 2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT) [Preprint]. doi:10.1109/ismsit.2019.8932845 (2019).

6. Parmar, M., et al.: Sentiment Analysis on interview transcripts: An application of NLP for quantitative analysis. In: 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI) [Preprint]. doi:10.1109/icacci.2018.8554498(2018).

7. Radha, N.: Video retrieval using speech and text in video. In: 2016 International Conference on Inventive Computation Technologies (ICICT) [Preprint]. doi:10.1109/inventive.2016.7824801 (2016).

8. Shubhang, S., et al.: Identification of hate speech and offensive content using bi-gru-lstm-cnn model. In: 2023 International Conference on Intelligent Data Communication Technologies andInternet of Things (IDCIoT) [Preprint]. doi:10.1109/idciot56793.2023.10053415 (2023).

9. Shukla, S., Nagpal, S., Sabharwal, S.: Hate speech detection in Hindi language using Bert and Convolution Neural Network. In: 2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS) [Preprint]. doi:10.1109/icccis56430.2022.10037649 (2022).

10. Solitana, N.T., Cheng, C.K.: Analyses of hate and non-hate expressions during election using NLP. In: 2021 International Conference on Asian Language Processing (IALP) [Preprint]. doi:10.1109/ialp54817.2021.9675186 (2021).