iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.51463

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

Abusive Language Detection in Speech Dataset

Anurag Rathore[1], Anmol Kumar[2], Aman Negi[3], Nidhi Chandra[4]
[1, 2, 3, 4]*Amity University , India*

Abstract: ***In the topic, speech recognition there is a rapid increasing in it into area of engineering technology. Speech recognition delivers many different kinds of pros and it's been uses in a multiple field. Having a different type of language placed a restriction of talking between people. From this project we will going to create and develop to supports different systems that allows persons in that place or situation to change of data through interacting with end device users by voice or speech, after developing this project we will destroy the barriers of communication. This project takes that into consideration and makes an attempt to guarantee that it can identify speech and transform audio input into text.***
The speech is converted into text format. To overcome the offensive content in real-time every social media platform should implement an effectual hate speech detection system. There are many ways from that we can classify hate speech such as Machine Learning, Rule Based, Deep Learning Based and Hybrid.
Keywords: Speech Recognition, Hate Speech, Offensive Language, Machine Learning, Rule Based, Deep Learning Based and Hybrid.

I. INTRODUCTION

Hate Speech recognition is proliferating(spreading) the data in the internet growing. So, we will do identification and do an investigation upon the problems that would be faced in auto-mated online tools or applications for the reason of text based hate speech identifications. The nuances of languages, different definitions of what constitute hate speech, and the limitations of the data present for the purpose of training and testing reasons these devices are some of the difficulties. In modern days many methods have interpretability issues, making difficulty for understanding why the system chooses a chosen course of action. . We suggest a multi-view SVM method that, while being simpler and giving more clearly understandable results than neural approaches.

Dismally, the hate crimes are not that new part to our culture. On social media and some others platforms that uses for virtual kind of communications are now playing an increasingly more precious role in field of hate crimes. Because of this, it has been suggested that social media plays a role in the radicalization of some suspects in recent terrorist incidents with a strong racial or religious undertone.

AS such recent terrorist attacks have suggested that social media or online platform for virtual communication been playing a role in the radicalization of some suspects with hard racial or religious affiliations. A one video that went viral in which this kind of incident happen in NZ and it was live streaming on face book platform.

The users can have a view or can express themselves independently and being anonymous on a extensive array of virtual communication platforms that includes Social media. The ability for express oneself independently is a human rights to be cherished, but inciting and spreading hate against another groups is an abuse for that freedoms.

The reason of this project was very important it's because this gives use people from different cultures, nationalities and languages so that they can share their own thoughts between among others. It removes the barriers among the people developed due to the language. Language translation system or technology has been brought us from being several nationalities together and has made great part of humanity in between people. It also simulates economic activity and has a vast impact on the society. Identifiaction of hate speech is a very difficult tasks and having some stages for that process. There are lot of unconvincing things in that data, so anything which is closest to that data need to be preprocessing. Various classifications of algorithms thenafter, recognize of abusive or hate speech in data. There are many different machine learning(ML) algorithms for detecting hate Speech, and also each algorithms is suitable for diverse scenarios

In this project we aim at developing an algorithm which would take various texts as input and check for the presence of hate speech in them. If there is any presence of hate speech in those text it would label that text as hate speech and if not then it would label it as a clean text. It would also be capable enough to identify the targeted victim of that hate speech (for ex. Hate against any of these caste, creed, gender, gender, nationality or religion). Although, while undergoing this project we discovered that hate speech is very indistinctive and has a lot of discriminative properties which makes it’s dataset to be difficult to detect in the long run.

©IJRASET: All Rights are Reserved

II. LITERATURE REVIEW

A lot of research has been conducted for the domain of video surveillance, object detection and machine learning. Some of the most notable work relevant to this project has been mentioned below

1) In 2014, C. J. Hutto et al. proposed an approach to classify sentiment using VADER, which is a rule-based approach [18]. They created a list of lexical features , After then they combined that list of lexical features with five general rules that encapsulate syntactical and grammatical rules for presenting sentiments. In 2015, They categorized the hate speech problem into three fields religion, nationality, and race. The main objective is to develop sentiment classification. he developed model not only detects subjective sentences but also classifies and ranks the polarity of sentiment phrases. Eventually, they achieved 71.5 % perfection using Lexion based approach.
2) Fatahillah et al. (2017) used Naive Bayes Classifier Algorithm to detect hate speech on Instagram using the k-nearest neighbor classifier [20]. They collected the data set using Twitter API from Twitter and annotated those data set manually. After preprocessing and feature engineering phase, they applied the Naive Bayes Classifier algorithm and found 93% of accuracy.
3) M. Ali Fauzi et al. (2018) proposed an approach to identify hate speech using a set of supervised learning algorithms [21]. They ensembled five different classification algorithms, including K-Nearest Neighbours, Random Forest, Naive Bayes, Support Vector Machine, and Maximum Entropy. They collected the data set using Twitter API and annotated those data set manually. In preprocessing phase, They employed tokenization, filtering, stemming, and term weighting methods. They utilized the bag of words features with TFIDF techniques. The naive Bayes algorithm performed best with 78.3 % of accuracy among all the other five stand-alone classifiers.
4) In 2019, P. Sari et al. proposed an approach to detect hate speech using logistic regression on Twitter. [22] They collected the data from Twitter and employed Case Folding, Tokenizing, Filtering, and Stemming methods in preprocessing phase. After Pre-processing, the TF-IDF technique is used for vectorization. After Feature engineering, the Logistic regression algorithm has been applied, and they have found 84% of accuracy.
5) Rui Zhao et al. (2015) proposed an approach to detect cyberbullying using Semantic-Enhanced Marginalized Denoising Auto-Encoder [24]. They used two sources of data set. The first source is Twitter, and the second source is Myspace. Twitter data was collected through Twitter stream API, and Myspace data was collected using the web crawling technique. They have achieved 84.9 % accuracy using smSDA for the Twitter dataset, and they have got 89.7% of accuracy with smSDA with the MySpace dataset
6) Axel Rodríguez et al. (2019) proposed an approach to detect hate speech content using sentiment analysis on Facebook [25]. They used Graph API to extract the post and comments from Facebook. To remove the unrelated texts VADER and JAMMIN were used. In preprocessing phase, they filtered out all unnecessary stopwords or symbols. Preprocessed documents converted into the vector using TFIDF. The resulting matrix is passed to the k-means clustering algorithm as an input matrix. The most negative articles and responses were collected using sentiment and emotion analysis.
7) Michele Di Capua et al. (2019) proposed an approach to detect cyberbullying using unsupervised learning [27]. They collected over 54,000 data set from YouTube and Annotated all data sets manually. The GHSOM network algorithm was implemented using the SOM-Toolbox-2 platform. They trained and tested GHSOM using a K-fold method with K = 10. As a result, they have got 64% of accuracy.
8) Tin Van Huynh et al. (2019) proposed an approach to detect hate speech using Bi-GRU-CNN-LSTM Model [29]. In this paper, they collected data from Twitter and categorized their data into three labels (OFFENSIVE, HATE, and CLEAN). After cleaning the data, they implemented three neural network models such as BiGRU-LSTM-CNN, Bi-GRU-CNN, and TextCNN to identify hate speech. They achieved a 70.57% of F1 score as a result.
9) Gambäck et al. (2019) utilized a deep learning algorithm to detect hate speech on Twitter [30]. In this paper, they collected data from Twitter and divided the data set into four categories(sexism, racism, combined(sexism and racism), and non-hate-speech). They employed four CNN models that were trained with character n-gram, word2vec, random vectors combined(word2vec and character n-gram). The author utilized a 10-fold technique to improve the accuracy of the model. Among all four models, word2vec based CNN model performed well with a 78.3% of F-score.
10) Viviana Patti et al. (2019) proposed a Hybrid based approach to detect hate speech [31]. In this paper, they employed two models. In their first model, they implemented a linear support vector classifier (LSVC), and in the second model, they employed a long short-term memory (LSTM) neural model with word embedding. They concatenated 17 categories, such as HurtLex, with two types, namely LSVC and LSTM. Joint learning with a multilingual word embedding model, including HurtLex, performed best with 68.7% of F1-score.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

11) There are many approaches by which hate speech detection can be carried out, such as Machine learning, Deep learning, and the Rule-based approach.

III. METHODOLOGIES

1) Phase 1

Using Twitter Dataset

The converted text is compared with the Twitter Dataset and then the output will be either hate speech or offensive language

2) Phase 2

Speech To Text Translation

With the help of various machine(ml) and deep learning(dl) technologies we will change the speech to the text. After that we will apply our hate speech detection algorithms to that text to detect the presence of hate speech in that data.

3) Phase 3

Data Preprocessing

In this phase duplicity of data and noise from the data is removed. Data cleaning is method in Data-mining which is to applied to remove the noisy data.

4) Phase 4

Count Vectorizer

Count Vectorizer is a great tool provided by the scikit-learn library in Python. It is used to transform a given text into a vector on the basis of the frequency (count) of each word that occurs in the entire text. This is helpful when we have multiple such texts, and we wish to convert each word in each text into vectors (for using in further text analysis).

5) Phase 5

Decision Tree Classifier

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

IV. RESULTS AND DISCUSSIONS

The Hate Speech detector is made with help of Twitter dataset and using this dataset we are able to check whether the speaker has used any hate speech offensive language.



```
,count,hate_speech,offensive_language,neither,class,tweet
0,3,0,0,3,2,!!! RT @mayasolovely: As a woman you shouldn't complain about cleaning up your house. &amp; as a man you should always take the trash out...
1,3,0,3,0,1,!!!!! RT @mleew17: boy dats cold...tyga dwn bad for cuffin dat hoe in the 1st place!!
2,3,0,3,0,1,!!!!!!! RT @UrKindofBrand Dawg!!!! RT @80sbaby4life: You ever fuck a bitch and she start to cry? You be confused as shit
3,3,0,2,1,1,!!!!!!!!! RT @C_G_Anderson: @viva_based she look like a tranny
4,6,0,6,0,1,!!!!!!!!!!!! RT @ShenikaRoberts: The shit you hear about me might be true or it might be faker than the bitch who told it to ya &#57361;
5,3,1,2,0,1,"!!!!!!!!!!!!!!!!!!""@T_Madison_x: The shit just blows me..claim you so faithful and down for somebody but still fucking with hoes! &#128514;&#128514;&#128514;"""
6,3,0,3,0,1,"!!!!!!""@__BrighterDays: I can not just sit up and HATE on another bitch .. I got too much shit going on!"""
7,3,0,3,0,1,!!!!&#8220;@selfiequeenbri: cause I'm tired of you big bitches coming for us skinny girls!!&#8221;
8,3,0,3,0,1,""" &amp; you might not get ya bitch back &amp; thats that """
9,3,1,2,0,1,""" @rhythmixx_ :hobbies include: fighting Mariam""

bitch"
10,3,0,3,0,1,""" Keeks is a bitch she curves everyone "" lol I walked into a conversation like this. Smh"
11,3,0,3,0,1,""" Murda Gang bitch its Gang Land """
12,3,0,2,1,1,""" So hoes that smoke are losers ? "" yea ... go on IG"
13,3,0,3,0,1,""" bad bitches is the only thing that i like """
14,3,1,2,0,1,""" bitch get up off me """
15,3,0,3,0,1,""" bitch nigga miss me with it """
16,3,0,3,0,1,""" bitch plz whatever """
17,3,1,2,0,1,""" bitch who do you love """
18,3,0,3,0,1,""" bitches get cut off everyday B """
19,3,0,3,0,1,""" black bottle &amp;amp; a bad bitch """
20,3,0,3,0,1,""" broke bitch cant tell me nothing """
21,3,0,3,0,1,""" cancel that bitch like Nino """
22,3,0,3,0,1,""" cant you see these hoes wont change """
23,3,0,3,0,1,""" fuck no that bitch dont even suck dick "" &#128514;&#128514;&#128514; the Kermit videos bout to fuck IG up"
24,3,0,3,0,1,""" got ya bitch tip toeing on my hardwood floors "" &#128514; http://t.co/cOU2WQ5L4q"
25,3,0,2,1,1,""" her pussy lips like Heaven doors "" &#128524;"
26,3,0,3,0,1,""" hoe what its hitting for """
27,3,0,3,0,1,""" i met that pussy on Ocean Dr . i gave that pussy a pill "" &#128524;"
28,3,0,3,0,1,""" i need a trippy bitch who fuck on Hennessy """
29,3,0,3,0,1,""" i spend my money how i want bitch its my business """
30,3,0,3,0,1,""" i txt my old bitch my new bitch pussy wetter """
```

Figure 1: The Twitter Dataset

In order to check through our detector, we need some important libraries and engines. We are using Jupyter we first convert our speech to text in order to do so we use two libraries

A. Speech Recognition

1) Recognize speech input from the microphone
2) Transcribe an audio file
3) Save audio data to an audio file
4) Show extended recognition results
5) Calibrate the recognizer energy threshold for ambient noise levels
6) Listening to a microphone in the background
7) Various other useful recognizer features

Pyttsx3 is a cross-platform speech (Mac OSX, Windows, and Linux) library. You can set voice metadata such as age, gender, id, language and name. The speech engine comes with a large amount of voices.

```
result2:
{   'alternative': [  {   'confidence': 0.87051958,
                          'transcript': 'I will kill you I will kill you'},
                      {'transcript': 'I will kill you'},
                      {'transcript': 'I will kill I will kill you'},
                      {'transcript': 'will kill you I will kill you'},
                      {'transcript': 'I will kill you will kill you'}],
    'final': True}
Did you say  i will kill you i will kill you
```

Figure 2: Speech to text Converted

Now the speech is converted to text we place the text to the sample and using twitter dataset the text is checked to whether be hate speech offensive language or normal text.

```
In [14]: #sample = "Let's unite and kill all the people who are protesting against the government"
         #sample = "you bitch"
         sample = "i kill you"
         #sample = "i will kill you "


         data = cv.transform([sample]).toarray()
         print(clf.predict(data))

         ['Hate Speech']
```

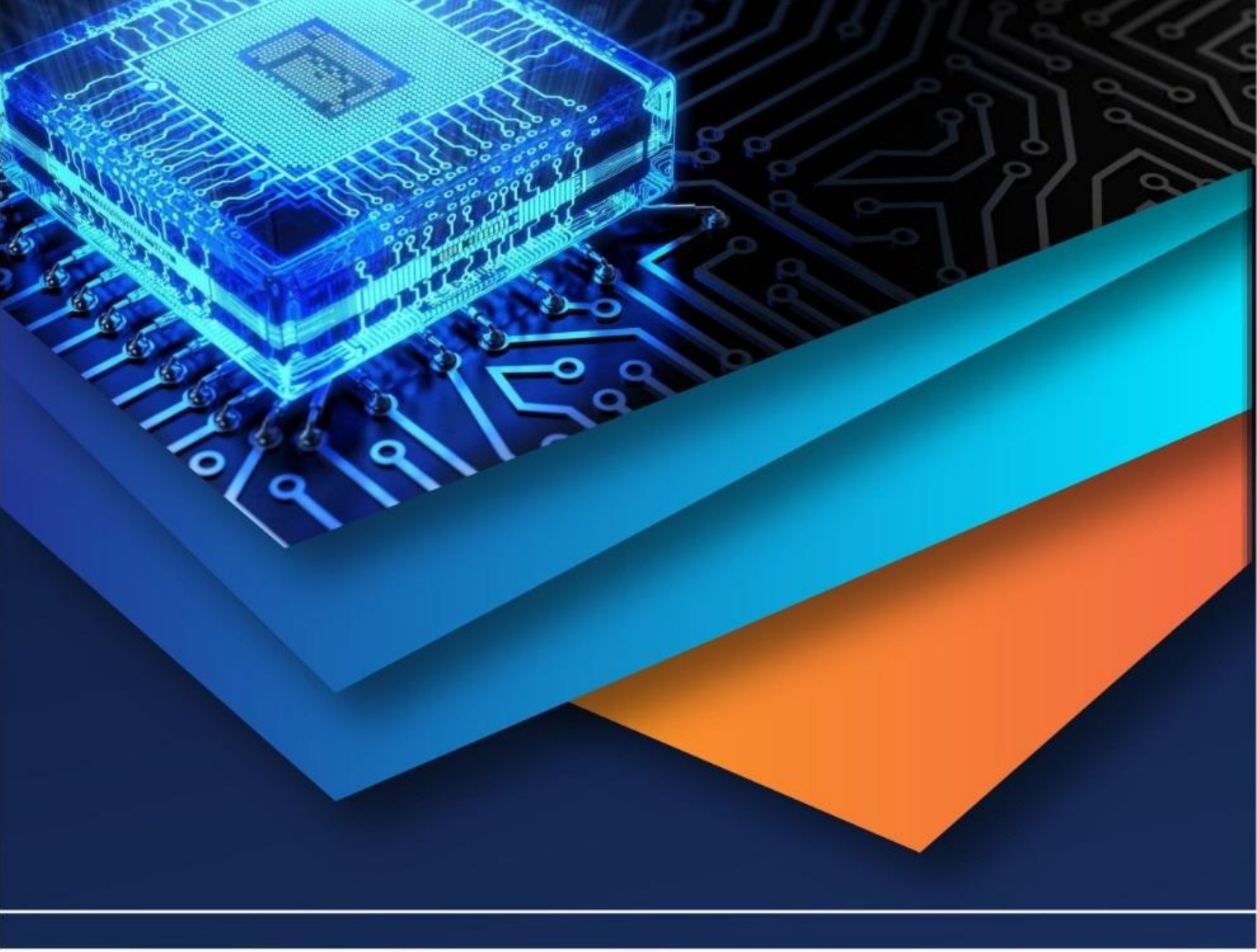
Figure 3: Hate Speech Detection

V. CONCLUSION

Due to the anonymity and mobility of such platforms as well as the shifting political landscape in many parts of the world, the spread of hate speech on social media has considerably increased in recent years. Despite significant effort by legislative authorities, law enforcement agencies, and social media firms, it is widely acknowledged that successful countermeasures rely on computerised semantic analysis of such information. The identification and classification of hate speech according to its targeting traits is a critical task in this direction and due to the numerous users who might compete with each other on Twitter, it is crucial for the success or ruin of one's image in today's social media. Examples of words with a negative connotation include those used in hate speech. Hate speech, which is included under Article 28 of the ITE Law, may be classified as having evil viewpoints. There are many people who, both knowingly and unknowingly, oppose hate speech on social media.

Social media, sadly, lacks the capability to compile data from a discourse already in progress into a conclusion.

Using text mining is one method for deriving conclusions from aggregate data. To categorise whether or not the sentence's text contains aspects of hate speech is the goal of this essay. In order for subsequent speech to be identified, the author of this research aims to develop a method for classifying hate speech elements in text using a computer. use the Multinomial Logistic Regression technique. The author thinks that after developing this programme, a computer will be able to detect and categorise hate speech in text posted on the social networking site Twitter. According to test findings, the average precision, recall, and accuracy were 80.02, 82%, and 87.68%, respectively.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)