Deepika Saraswat,2023, 11:6
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology

An Open Access Journal

# Twitter Fake Speech Detection Using Model of Machine Learning and NLP

**Deepika Saraswat, Dr. Nirupama Tiwari**
Advance Computing Department
Sage University, Indore,MP,India

**Abstract-Social Networking platforms are the most efficient way to express or convey one's feelings or thoughts. The growth of social media platforms has led people to indulge in illegal and unethical activities. People started using social media platforms as a tool to express their hate, anger, and criticism towards an individual or an ethnic group. Today the use of Twitter is increasing day by day and the fact that most people come here to express their thoughts regarding social or economic problems, but some people use Twitter as a platform to target and spread hate towards someone based on sex, religion, race, etc using hateful hash tags. In this paper, we will be using Twitter tweets and NLP sentiment analysis techniques to detect whether the tweet is hateful or not. By categorizing the tweets into the label 0 and 1, where 0 represents non-hateful speech and 1 represents hateful speech. This helps us to detect and control hate speech.**

**Keywords-Sentiment Analysis, ClassificationTechniques, NLP,Datamining, Twitter Dataset.**

## I. INTRODUCTION

In Today's Era there is an enormous amount of data that is continuously increasing day by day even seconds by second. Social Media has become a former medium to analyze mob feedback, review, and opinion about a certain technology, economic growth, and social welfare. It has become a vital place to analyze –What Is the opinion of the crowd, Being a platform used by millions of users it providesan efficient way to know about the crowd. In this paper, we present a solution to detect hateful thoughts or hate speech towards an individual or a group of people based on their religion, caste, race, sex, etc

### 1. Hate Speech

The term "hate speech" refers to partial, aggressive, and malicious speech that targets an individual or a group of people because of their conscious or unconscious intrinsic characteristics. There is no specific definition of hate speech under international human rights law. The concept of hate speech is still widely disputed especially about the right to free speech.

1.According to United Nation Strategy and Plan of Action on Hate Speech defines hate speech as"any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor."

While the above is not a legal definition and is broader than "incitement to discrimination, hostility or violence" – which is prohibited under international human rights law -- it has three important attributes:

- Hate speech can be conveyed through any form of expression,including images, cartoons, memes, objects, gestures and symbols and it can be disseminated offline or online.
- Hate speech is "discriminatory" (biased, bigoted or intolerant) or "pejorative" (prejudiced, contemptuous or demeaning) of an individual or group.

- Hate speech calls out real or perceived "identity factors" of an individual or a group including: "religion, ethnicity, nationality, race, colour, descent, gender," but also characteristics such as language, economic or social origin,
- disability, health status, or sexual orientation, among many others

**2. Sentiment Analysis**

The sentiment is an NLP technique to determine whether the given data(textual) is positive, negative, or neutral. Analysis of sentiment can be viewed as a way of evaluating people for particular incidents, labels, goods, or businesses. The sentiments are categorized among positive and negative sentiments. In this analysis, we have used Python to apply a classification algorithm and to perform sentiment analysis using Natural Language processing techniques.

The end goal of this analysis is to apply sentiment analysis to the collected tweets so that it can be determined if the tweets can be considered hateful or not. We also want to use sentiment analysis to identify tweets. that express an opinion (subjective tweets) as compared to those that just provide information without a positive or negative opinion.

2. The first step of building our model was to balance the number of hate and non-hate tweets. Our data preprocessing step involved 2 approaches, Bag of words and Term Frequency Inverse Document Frequency (TFIDF).

The bag-of-words approach is a simplified representation used in natural language processing and information retrieval. In this approach, a text such as a sentence or a document is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity. TFIDF is a numerical statistic that is intended to reflect how important a word is to a document in a collection. It is used as a weighting factor in searches for information retrieval, text mining, and user modeling. Before we input this data into various algorithms, we have to clean it as the tweets contain many different tenses, grammatical errors, unknown symbols, hashtags, and Greek characters. We tackle this problem by employing lemmatization, stemming, removal of stop words, and omissions. Lemmatization removes the inflectional endings of words and returns the word to its base or dictionary form of it. Stemming is similar to lemmatization in that it reduces the inflected or derived words to their word stem. A stop word is a commonly used word, such as "the", "a", "an", and "in", that we are programmed to ignore as it holds no importance. The last step is to omit any foreign characters and Greek symbols

## II. REVIEW OF LITERATURE

The paper research that we present takes account of previousstudies in the issue field of sentiment analysis on hate speech on social media.Ricardo Martins, Marco Gomes, Jose Joao Almeida, Paulo Novais, Pedro Henriques [3] Presented a combination of lexicon based and machine learning approaches to predict hate speech contained in a text, using an emotional approach through sentiment analysis.

In 2013, N. Sambuli et al. worked on a project called "Umati Monitoring Online Dangerous Speech." The project was based on monitoring Hatebase and dangerous speech[4]. According to them, dangerous expressions can be observed in the following ways:

1. It is targeted to a group of people and not a single person. Dangerous speech is an offensive speech that encourages the audience to participate in acts of violence against a particular group of people, therefore In the internet domain, the most prevalent forms of hate speech are related to religion, race, sexual orientation, nationality, class, and gender.
2. Hate Speech may contain one of the pillars of dangerous speech, for instance, statements that classify people as vermin, which claims that a group of people is like rodents or insects.
3. Dangerous speech often incites the listener to support or commit acts of violence against the specific group. The six most common calls

action in dangerous speech are: kill, riot, beat, loot, forcefully evict, and discrimination. In another report, VedantK shirsagar, Adwait Toro, Sushrut Shendre, Sneha Lakshmi kanthaiah, Rishab Ballekere Narayana Gowda, Tanvir Brar, Chavi Singal. Detecting Hate tweets — Twitter Sentiment Analysis[8] Proposed a solution to the detection of hate speech and offensive language on Twitter through machine learning using Bag of Words and TF IDF values. We performed

comparative analysis of Logistic Regression, Naive Bayes, Decision Tree, Random Forest and Gradient Boosting on various sets of feature values and model parameters. The results showed that Logistic Regression performs comparatively better

## III. PROPOSED METHODOLOGY

### 1. Data Collection

The very first step was data collection.We have used a data scrapping tool named as Snscrape.[5] Snscrape is a scraping tool for social networking services(SNS). It scrapes information like user profiles, hashtags, searches, and threads and returns the discovered items, e.g. the relevant posts. It was released on July 8,2020 and it is capable of scraping data from a variety of platforms for example Twitter, Instagram, Reddit, Facebook, Telegram etc. It requires python 3.8 or higher. By using inscape we can scrape the data by running a query in which we can search data on the basis username, hashtags, specific keywords, etc. We can also set the limit of the data gathering as per requirement, using this tool helps us to collect a datain specific time period.

```
import snscrape.modules.twitter as sntwitter
import pandas as pd

query = "#kill until:2022-10-26 since:2022-06-01"
tweets = []
limit = 300


for tweet in sntwitter.TwitterSearchScraper(query).get_items():

    # print(vars(tweet))
    # break
    if len(tweets) == limit:
        break
    else:
        tweets.append([tweet.date, tweet.username, tweet.content])

df = pd.DataFrame(tweets, columns=['Date', 'User', 'Tweet'])
print(df)

# to save to csv
df.to_csv('train.csv')
from google.colab import files
files.download("train.csv")
```

Figure 1. Snscrape Scrapping Tool.

```
                  Date             User  \
0   2022-10-25 23:24:51+00:00     snoopbee1
1   2022-10-25 23:18:05+00:00    news__reddit
2   2022-10-25 23:08:19+00:00  NarratamiSempai
3   2022-10-25 22:25:03+00:00        IamAbot94
4   2022-10-25 22:24:51+00:00     snoopbee1
..                        ...           ...
295 2022-10-21 02:04:53+00:00     snoopbee1
296 2022-10-21 01:56:13+00:00     trevystan
297 2022-10-21 00:49:02+00:00     noa_order
298 2022-10-21 00:12:27+00:00  SnopesRatingBot
299 2022-10-20 23:54:33+00:00     magbeigames

                                                 Tweet
0    #SaturdayWisdom  5 ways to #kill #cellulite at...
1    Police kill two dogs after US Amazon driver di...
2    Is this why #parents #kill themselves and/or #...
3    #SaturdayThoughts  5 ways to #kill #cellulite ...
..                                                 ...
295  #DailyThoughts  5 ways to #kill #cellulite at ...
296                                  #kill #donaldson
297  #HORROR COMES TO #RUNCORN\nTHIS #HALLOWEEN!\n\...
298  ✂ False\n#factcheck #lion #trainer #mgm #kill\...
299  Play Bill The Bowman game online for FREE! An ...

[300 rows x 3 columns]
```

Figure 2. Sample Output

| | Date | User | Tweet |
|---|---|---|---|
| 0 | 2022-10-2 | snoopbee1 | #SaturdayWisdom 5 ways to #kill #cellulite at https://t.co/4NdPan |
| 1 | 2022-10-2 | news__red | Police kill |
| 2 | 2022-10-2 | Narratami | Is this why #parents #kill themselves and/or #raise their #kids to no |
| 3 | 2022-10-2 | IamAbot94 | Please |
| 4 | 2022-10-2 | snoopbee1 | #SaturdayThoughts 5 ways to #kill #cellulite at https://t.co/4NdPa |
| 5 | 2022-10-2 | DebraLGril | @american2084 No need for assault #weapons in #America. Only t |
| 6 | 2022-10-2 | Scott_A_N | Crosshair placement finally getting better - Double kill #Valorant #K |
| 7 | 2022-10-2 | Scott_A_N | 3 Headshots and an assist, trying to work on my movement and pos |
| 8 | 2022-10-2 | BodySnatc | ðŸ©¸ |
| 9 | 2022-10-2 | BodySnatc | ðŸ©¸ |
| 10 | 2022-10-2 | joshuaalee | Neon CLUTCH 4K Neon CLUTCH 4K #Valorant #Kill âžĩ https://t.c |
| 11 | 2022-10-2 | forgerelli | Listen to |
| 12 | 2022-10-2 | efret25 | Red Team Fundamentals - I have just completed this room! Check i |
| 13 | 2022-10-2 | Narratami | There, I said it before you set off a stupid #Nuke to #kill everything |

Figure 3. Representation Of Sample Dataset In .Csv Format

### 2. Importing Libraries

After analyzing the data our next step is to import the required libraries for our project. Some of the libraries we use in this project are pandas, numpy, scikit learn, and nltk.

```
import pandas as pd
import numpy as np
import re
import seaborn as sns
import matplotlib.pyplot as plt
from matplotlib import style
style.use('ggplot')
from google.colab import files
import io
import nltk
nltk.download('punkt')
nltk.download('wordnet')
nltk.download('omw-1.4')
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer
from nltk.corpus import stopwords
nltk.download('stopwords')
stop_words = set(stopwords.words('english'))
from wordcloud import WordCloud
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix, ConfusionMa
```

Figure 4. Importing Libraries

### 3. Data Preprocessing

While tweets are collected in real time, Data cleaning was important , the following procedure is carried out.

- Firstly Remove all non-alphabetic characters.
- Remove duplicates if any.
- We also converted Apostrophe‟s to their respective complete sentences. In order to get complete meaning of the sentence So we can reach the proper context of the sentence.
- Since emoji‟s plays an important role in reflecting someone‟s emotions or feeling. We also converted emoji‟s to their respective meaning.
- Stop words are subsequently removed from tweets based on membership in the “stop words” corpus of the Natural Language Toolkit but keeping some important words like “ not ” .
- Then Stemming and Lemmatization is done.
-

**Emoji Conversion**

```
# Emotion detection by different symbols
emotion_dict = {
":)": "happy",
":-)": "happy",
":-]": "happy",
":-3": "happy",
":->": "happy",
"8-)": "happy",
":-}": "happy",
":o)": "happy",
":c)": "happy",
":^)": "happy",
"=]": "happy",
"=)": "happy",
"<3": "happy",
":-(": "sad",
":(": "sad",
":c": "sad",
":<": "sad",
```

**Data Cleaning**

```
# Apostrophe Dictionary
apostrophe_dict = {
"ain't": "am not / are not",
"aren't": "are not / am not",
"can't": "cannot",
"can't've": "cannot have",
"'cause": "because",
"could've": "could have",
"couldn't": "could not",
"couldn't've": "could not have",
"didn't": "did not",
"doesn't": "does not",
"don't": "do not",
"hadn't": "had not",
"hadn't've": "had not have",
"hasn't": "has not",
"haven't": "have not",
"he'd": "he had / he would",
"he'd've": "he would have",
"he'll": "he shall / he will",
"he'll've": "he shall have / he will have",
```

Figure 5. Emoji CoversionAnd Data Cleaning

From The Image Below The Difference Between Original Tweets And The Cleaned Tweets Can Be Seen.

| tweet |
| --- |
| ate @user isz that youuu?ð□□□ð□□□ð□□□ð□□□ð□□□ð... |
| to see nina turner on the airwaves trying to... |
| listening to sad songs on a monday morning otw... |
| @user #sikh #temple vandalised in in #calgary,... |
| thank you @user for you follow |

Figure 6. Original Tweets

| tweet |
| --- |
| ate user isz youuuâï |
| see nina turner airwaves trying wrap mantle ge... |
| listening sad songs monday morning otw work sad |
| user sikh temple vandalised calgary wso condem... |
| thank user follow |

Figure 7. Clean Tweets

### 4. Data Visualization

Word Cloud: [6]Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. Significant textual data points can be highlighted using a word cloud. Word clouds are widely used for analyzing data from social network websites. For generating word cloud in Python, modules needed are – matplotlib, pandas and wordcloud.

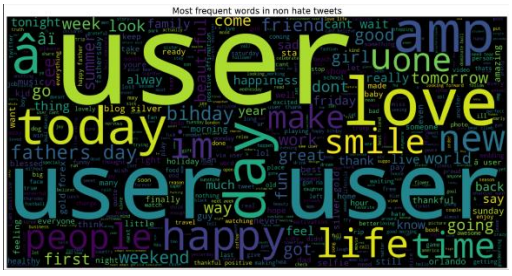


Figure 9. Word Cloud for non hate tweets

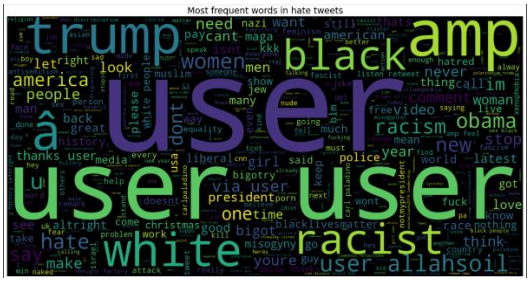


Figure 10. Word Cloud for hate tweets

## IV. MODEL BUILDING

The model procedure followed fundamental machine learning protocol, the overall dataset was split into two sections, 80 percent for training the model and 20 percent for testing it. As we have the historic data with the target variable therefore we will go with supervised machine learning approach. Also our target variable an only attain two values i.e 0 and 1 where 1 stand for hate speech and 0 stand for non hate speech hence our target variable is a categorical field. We tried using Logistic Regression, Random Forest and Decision Tree Classifier . [7]Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems.On the other hand Random Forest is another tree based algorithm It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. Logistic regression unlike the name suggest is an classification algorithm It is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set.

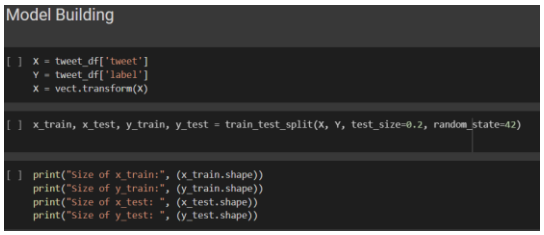


Figure 11. Model Building

## V. RESULT

### 1.Decision Tree

```
from sklearn.tree import DecisionTreeClassifier
clf = DecisionTreeClassifier()
clf.fit(x_train, y_train)
clf_predict = clf.predict(x_test)
clf_acc = accuracy_score(clf_predict, y_test)
print("Test accuarcy: {:.2f}%".format(clf_acc*100))
```

Figure 11. Decision Tree

The overall Test Accuracy from Decision Tree was approx 82%.

### 2. Logistic Regression

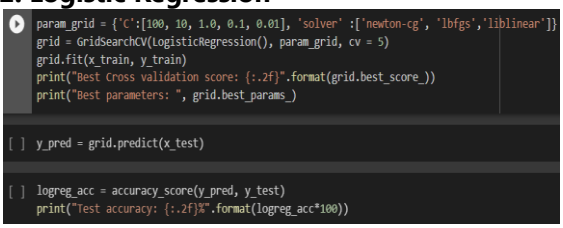


Figure 12. Logistic Regression

The overall Test Accuracy From Logistic Regression was approx 80%.

### 3.Random Forest

```
from sklearn.ensemble import RandomForestClassifier
clf = RandomForestClassifier(n_estimators = 100)
clf.fit(x_train, y_train)
clf_predict = clf.predict(x_test)
clf_acc = accuracy_score(clf_predict, y_test)
print("Test accuarcy: {:.2f}%".format(clf_acc*100))
```

Figure 13. Random Forest.

The overall Test Accuracy From Random Forest was approx 85%.

## VI. CONCLUSION

We are able to detect the hate speech to a certain extent using NLP and Machine Learning classifiers like Logistic Regression, Decision Tree, Random Forest Classifier.The best classifier for our model was Random Forest Classifier with an accuracy of approx 85%. Further For some Complex dataset our system fails to do so. It may be taken as future scope of Research

## REFERENCES

1. https:/ /hate speech/understanding-hate-speech/what-is-hate-speech
2. Vedant Kshirsagar, Adwait Toro, Sushrut Shendre, Sneha Lakshmi kanthaiah, Rishab Ballekere NarayanaGowda, TanvirBrar, ChaviSingal.Detecting Hate tweets — Twitter Sentiment Analysis. Available at:https://towardsdatascience.com/detecting-hate-tweets-twitter-sentiment-analysis-780d8a82d4f6.
3. Ricardo Martins, Marco Gomes, Jose Joao Almeida, Paulo Novais, Pedro Henriques."Hate speech classification in social media using emotional analysis".DOI:10.1109/BRACIS.2018.00019
4. N. Sambuli, F. Morara, & C. Mahihu. (2013). Umati: Monitoring online dangerous speech. Available at: http://www.ihub.co.ke/og/wpcontent/uploads/2014/06/2013-report-1.pdf.
5. https://github.com/JustAnotherArchivist/snscrape
6. https://www.geeksforgeeks.org/generating-word-cloud-python/
7. Carolina Bento "Decision Tree Classifier explained in real-life: picking a vacation destination". Available at: https://towardsdatascience.com/decision-tree-classifier-explained-in-real-life-picking-a-vacation-destination-6226b2b60575

**Authors Profile**

Prof. DeepikaSaraswat as an Asst. Professor in Advance Computing Department in SAGE UNIVERSITY, INDORE and having 16 years of Academic Experience. Pursuing PhD in Computer Science & Engineering. Received MTech in Computer Science & Engineering from ITM University, Gwalior. I has guided several students at Under Graduate level. My areas of current research include Machine Learning, Machine Learning and Deep Learning. I has published more than 10 research papers in the journals and conferences of international repute. I have the memberships of various Academic/ Scientific societies.

DR.NIRUPMA TIWARI as an Associate Professor in IAC Department in SAGE UNIVERSITY INDORE and having 15 years of Academic and Professional experience. She received PhD in Computer Science Engineering from Suresh GyanVihar University Jaipur India. M.tech degree in Software System from SATI Vidisha. She has guided several students at Master and Under Graduate level. Her areas of current research include Data mining and Image Processing Machine Learning and Deep Learning. She has published more than 30 research papers in the journals and conferences of international repute. She is having the memberships of various Academic/ Scientific societies**.**