International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 4, Issue 12, ISSN No. 2455-2143, Pages 558-561
Published Online April 2020 in IJEAST (http://www.ijeast.com)

HATE SPEECH DETECTION IN TWEETS USING MACHINE LEARNING ALGORITHM

Mrs. KopparthiHarika
Pragati Engineering College,
Kakinada, Andhra Pradesh, India

CH Mounika, T Anuradha, P Sharon,
B.Tech, Computer Science and Engineering,
Pragati Engineering College,
Kakinada, Andhra Pradesh, India,

Abstract— **Hate speech is currently of broad and current interest in the domain of social media. The anonymity and flexibility afforded by the Internet has made it easy for users to communicate in an aggressive manner. And as the amount of online hate speech is increasing, methods that automatically detect hate speech is very much required. Moreover, these problems have also been attracting the Natural Language Processing and Machine Learning communities a lot. Therefore, the goal of this paper is to look at how Natural Language Processing applies in detecting hate speech. Furthermore, this paper also applies a current technique in this field on a dataset.**

I. INTRODUCTION

Natural Language Processing (NLP) of NLP is sentiment analysis. From opinion polls to creating entire marketing strategies, this domain has completely reshaped the way businesses work, which is why this is an area every data scientist must be familiar with. In order to create a hate-speech-detecting algorithm, we are going to use Python-based NLP machine learning techniques. Machine learning is basically teaching machines to accomplish various tasks by training them through data. In this case, we are going to collect data from the Korean radical anti-male website, Womad, but you are free to use different kinds of data as long as the data is labeled appropriately (more on that later). Then, using a NLP (or Natural Language Processing) technique called Tf-Idf vectorization, well extract keywords that convey importance within hate speech.

II. PROPOSED ALGORITHM

Methodology of sentimental analysis in twitter mainly involves 5 steps.

2.1 DATA COLLECTION/ TWEET EXTRACTION

	id	label	tweet
0	1	0	@user when a father is dysfunctional and is so selfish he drags his kids into his dysfunction. #run
1	2	0	@user @user thanks for #lyft credit i can't use cause they don't offer wheelchair vans in pdx. #disapointed #getthanked
2	3	0	bihday your majesty
3	4	0	#model i love u take with u all the time in ur[illegible]
4	5	0	factsguide: society now #motivation

Fig. Tweets Preprocessing and cleaning

Twitter API — A Python wrapper for performing API requests. For fetching the twitter data from the twitter API includes the following steps 1] Installation of the needed software 2] authentication of twitters data. The main installation softwares include tweepy, text blob, nltk etc, Authentication involves different stepsstep1: visit the twitter website and click the button ‘ create new app’ . Step2:fill the details in the form provided and submit.Step3:It will be redirected to the app page where the “ ‘ consumer keys’ , ‘ consumer access’ , ‘ access token’ and ‘ access token secret’ “ that is needed to access the twitter data will be present.Step4:implement in python. There are different sources for storing the data taken from the twitter. They are like MongoDB, open source document storage database and is the go-to “ No SQL” database. It makes working with a database feel like working with JavaScript. PyMongo, a Python wrapper for interfacing with a MongoDB instance. This library lets you connect your Python scripts with your database and read/insert records. This is an example of the data that is been extracted from the twitter on the topic.

2.2 PRE-PROCESSING

Once the data is collected from the twitter the next step is preprocessing that is implemented in python. There are several steps involved in the preprocessing stage. They are,

1. Converting all uppercase letters to lowercase.
2. Tokenization generally done by installing the NLP package. It generally means removal of hash tags, numbers (1, 2, 3 etc.,), URL’ s and targets (@). Once

tokenization is over we move to the next step of preprocessing.

```
tokenized_tweet = combi['tidy_tweet'].apply(lambda x: x.split())
tokenized_tweet.head()
```

```
0                 [when, father, dysfunctional, selfish, drags, kids,
1     [thanks, #lyft, credit, cause, they, offer, wheelchair, vans, #d
2
3                                                              [#model,
4                                                            [factsguid
Name: tidy_tweet, dtype: object
```

3. Removal of non-English words

Twitter generally supports more than 60 languages. But our project mainly involves English tweets; hence we remove the non-English words.

	id	label	tweet	
0	1	0.0	@user when a father is dysfunctional and is so selfish he drags his kids into his dysfunction. #run	when fath
1	2	0.0	@user @user thanks for #lyft credit i can't use cause they don't offer wheelchair vans in pdx. #disapointed #getthanked	thanks #
2	3	0.0	bihday your majesty	
3	4	0.0	#model i love u take with u all the time in urð□□±!!! ð□□□ð□□□ð□□□ð□□□ ð□□¦ð□□¦ð□□¦	
4	5	0.0	factsguide: society now #motivation	

4. Emoticon replacements Emoticons are very important in determining the sentiment. So the emoticons are replaced by their polarity by seeing the emoticon dictionary.

```
0              [when, father, dysfunct, selfish, drag, kid,
1    [thank, #lyft, credit, caus, they, offer, wheelchair, van,
2                                                            [
3                                                   [#model, l
4                                                       [facts
Name: tidy_tweet, dtype: object
```

2.3 FEATURE EXTRACTIONS

election of useful words from the tweet is called as feature extraction. In the feature extraction method, we extract the aspects from the pre-processed twitter
dataset.
1. There are three different types of features namely unigram, bigram, n-gram
features.
2. Parts Of Speech Tags such as like adjectives, adverbs, verbs and nouns are good indicators of subjectivity and sentiment.
3. Negation is another important but difficult feature to interpret. The presence of a negation usually changes the polarity of the sentiment.

2.4 FEATURE SELECTION

Correct feature selection techniques are used in sentiment analysis that has got a significant role for identifying relevant attributes and increasing classification (machine learning) accuracy. They are categorized into 4 main types namely,1. Natural language processing 2. Statistical

2.5 ADVANTAGES OF NEURAL NETWORKS

The main advantage is that they are data driven self-adaptive methods where, they can adjust themselves to the data without any explicit specifications of functional or distributional form. They can approximate any function with arbitrary accuracy since they are universal functional approximates.

III. EXECUTION AND RESULT

Step 1: Open the source code in the Pycharm platform

```
import numpy as np
import pandas as pd
import os
print(os.listdir("../input"))
train = pd.read_csy("../input/train_E6oV3lV.csv")
test = pd.read_csv("../input/test_tweets_anuFYb8.
train.head()
test.head()
train['label'] = train['label'].astype('category'
train.info()
from nltk.stem import WordNetLemmatizer
from nltk import tokenize
from sklearn.feature_extraction.text import Tfidf
import re
train['text_lem'] = [''.join([WordNetLemmatizer()
    -z]',' ',text)) for text in lis]) for lis in
test['text_lem'] = [''.join([WordNetLemmatizer().
    ,' ',text)) for text in lis]) for lis in test
    .model_selection import train_test_split

X_train,X_test,y_train,y_test = train_test_split(
    train['label'])
```

Step 2:Tokenization

```
Tweet Tokenization

0    [when, father, dysfunctional, selfish, drags,'...
1    [thanks, #lyft, credit, can't, cause, they, do...
2                             [bihday, your, majesty]
3    [#model, love, take, with, time, urð±!!!, ð±!!!, ð
```

Step 3: Run Code

A. Understanding the common words used in the tweets

Now I want to see how well the given sentiments are distributed across the train dataset. One way to accomplish this task is by understanding the common words by plotting wordclouds.

A wordcloud is a visualization wherein the most frequent words appear in large size and the less frequent words appear in smaller sizes.

Lets visualize all the words our data using the wordcloud plot.

International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 4, Issue 12, ISSN No. 2455-2143, Pages 558-561
Published Online April 2020 in IJEAST (http://www.ijeast.com)

Fig. Understanding the common words used in the tweets

We can see most of the words are positive or neutral. With happy and love being the most frequent ones. It doesn't give us any idea about the words associated with the racist/sexist tweets. Hence, we will plot separate wordclouds for both the classes (racist/sexist or not) in our train data.

B. Words in non racist/sexist tweets

Fig 4.2.5 racist/sexist tweets

We can see most of the words are positive or neutral. With happy, smile, and love being the most frequent ones. Hence, most of the frequent words are compatible with the sentiment which is non racist/sexists tweets. Similarly, we will plot the word cloud for the other sentiment. Expect to see negative, racist, and sexist terms.

C. Racist/Sexist Tweets

Fig 4.2.4 words in non racist/sexist tweets

As we can clearly see, most of the words have negative connotations. So, it seems we have a pretty good text data to work on. Next we will the hashtags/trends in our twitter data.

D. Non-Racist/Sexist Tweets

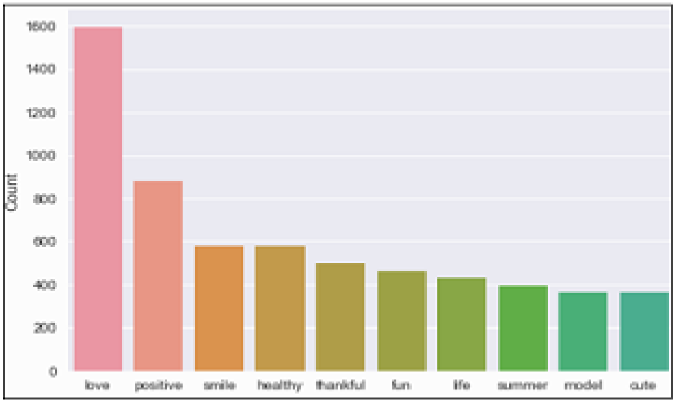

Fig 4.2.6 hashtags appearing in the non racist/sexist tweets.

All these hashtags are positive and it makes sense. I am expecting negative terms in the plot of the second list. Let's check the most frequent hashtags appearing in the racist/sexist tweets.

E. Racist/Sexist Tweets

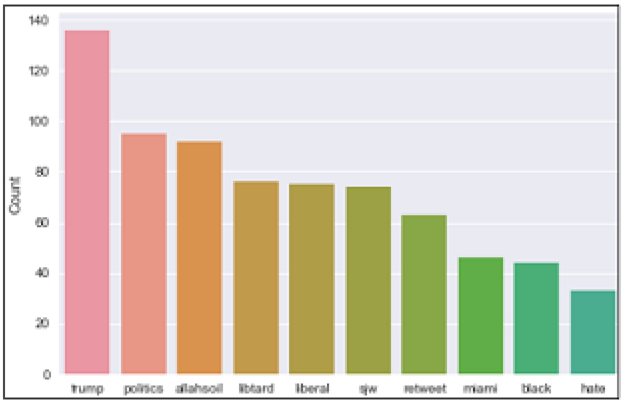

Fig 4.2.7 Negative Racist/Sexist Tweets

As expected, most of the terms are negative with a few neutral terms as well. So, it's not a bad idea to keep these hashtags in our data as they contain useful information. Next, we will try to extract features from the tokenized tweets.

IV. CONCLUSION

In this paper, the aim was to detect hate speech using a Natural Language Processing technique. To enable successful execution of the research it was first necessary to understand what hate speech is. To accomplish this, an overview of this topic has been conducted. Here it can be concluded that hate speech has several definitions, all coming from different platforms. Hate speech detection is a classification-related task, and that's why further literature was reviewed to understand the idea behind Natural Language Processing and the application of various techniques. Previous work showed that deep learning models improve the state-of-art approaches within hate speech classification tasks. Therefore, a deep learning method, namely a Convolutional Neural Network (CNN), has been applied on a Twitter dataset. This data contains tweets annotated with three labels: hate, offensive language and neither.

V. REFERENCE

1) M. Ott, Y. Choi, "Finding deceptive opinion spam by any stretch of the imagination; Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume; pp. 309–319, 2011;
2) C. Cardie , J.T. Hancock Negative Deceptive Opinion Spam; Proceedings of the HLT-NAACL; pp. 497–501, 2013.
3) C.H. Yu , M.W. Ward, M. Morabito, W. Ding, Crime forecasting using data mining techniques; Proceedings of the International Conference on Data Mining, ICDM'11; Vancouver, BC, CanadaCanada, pp. 779–789, 2011;
4) A. Bermingham, A. Smeaton, On using Twitter to monitor political sentiment and predict election results; Proceedings of the Workshop on Sentiment Analysis Where AI meets Psychology (SAAIP 2011); Chiang Mai, Thailand, pp. 2–10, 2011;
5) X. Wang , D. W. Brown , M. S. Gerber, Spatio-temporal modeling of criminal incidents using geographic, demographic, and Twitter-derived information; Proceedings of the 2012 IEEE International Conference on Intelligence and Security Informatics (ISI); Arlington, pp. 36–41, 2012;
6) J. Bollen , H. Mao , X. Zeng, Twitter mood predicts the stock market. J. Comput, pp. 1–8, 2011;
7) P.T. Metaxas , E Mustafaraj , D. Gayo-AvelloGayo-Avello, How (not) to predict elections; Proceedings of the 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing; Boston, pp. 165–171, 2011;
8) M. Ott, C. Cardie , J. K. Hancock, Negative Deceptive Opinion Spam; Proceedings of the HLT-NAACL, pp. 497–501, 2013;
9) A.K. Uysal , S. Gunal, A novel probabilistic feature selection method for text classification. Knowl.-Based Syst, pp. 226–235, 2012;
10) D.M.W. Powers, Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation, pp. 37–63, 2011;
11) M.L. McHugh, Interrater reliability: The kappa statistic. Biochem. Med. Biochem. Med. 22:276–282. doi: 10.11613/BM.2012.031, 2012;
12) J. Heidemann, M. Klier, F. Probst, Online social networks: A survey of a global phenomenon.doi: 10.1016/j.comnet.2012
13) D. Choudhury, A. Paul, Community detection in social networks: An Over, pp. 6–13, 2013;

VI. ACKNOWLEDGMENT

It gives us an immense pleasure to express deep sense of gratitude to our guide Mrs. Kopparthi Harika, Department of Computer Science & Engineering because of his whole hearted and invaluable guidance throughout the report. Without his sustained and sincere effort, this report would not have taken this shape. She encouraged and helped us to overcome various difficulties that we have faced at various stages of our report.
We would like to sincerely thank Dr. M. Radhika Mani, Professor & HOD,
Computer Science & Engineering, for providing all the necessary facilities that led to the successful completion of our report.
We would like to take this opportunity to thank our beloved Principal and Vice
Principal Dr. S. Sambhu Prasad, Dr. K. Satyanarayana for providing a great
support to us in completing our project and for giving us the opportunity of doing the mini project report.