PREDICT HATE TWEETS USING LOGISTIC REGRESSION. DATA MINING

Table of Contents

1. Executive Summary:

This report encapsulates key findings of classifying sentiments of a bunch of tweets obtained from Twitter. The reason to choose this topic was to summarize the nature of reactions being posted on Twitter and utilizing the methodology described in this report; users with consistent record of posting negative tweets could be monitored closely so that hateful conduct or violence is not promoted. The primary components of the analysis include 1) calculation of sentimental labels 2) data preprocessing and cleansing 3) statistical model for classification 4) critical analysis & summary of findings.

2. Problem Statement:

We live in an era, where the internet is easily accessible. With the internet came a lot of social platforms that connect people albeit their physical distances. These social platforms did not just connect people but boosted the e-commerce sector. With the social platforms, adding values to people's lives it comes with some curses. Social platforms, be it Twitter or Instagram, have a lot of contents that can be labelled as derogatory, insensitive and malicious. This has eventually increased the rate of hate crimes drastically. It is imperative to guard social platforms against these crimes. Because of the volatility and the amount of data on these platforms, it is impossible to manage the contents on these social platforms manually.

Twitter is one of the most prominent social platforms. But at the same time there has been a lot of hate crimes occurring over twitter. The minimum age eligibility to hold a twitter account is thirteen years. The early teens may fall victims of mental stress due to the negative tweets posted on Twitter.

The goal is to build a model that can predict without any bias a tweet is negative or positive.

2.1 Data Description:

Data Source: Data is read into a data frame using code mentioned in 7.2.1https://www.kaggle.com/vkrahul/twitter-hate-speech

2.2 Data Dictionary:

Field Name	Description
id	Serial Number
tweet	The twitter tweets
Sentiment	Label of the tweet 1 – Negative tweet 0 – Positive tweet

The project analyzes a dataset from Kaggle, "Twitter hate speech" consisting of about 32k tweets. It consists of tweets twitted by users in the most popular social media platform twitter categorized as hate or not. We tried to think out of the box and generated our own labels for each tweet using Sentiment Analysis.

After sentiment analysis on these tweets, the new target labels are generated by calculating the positive and negative sentiments of each tweet and classifying them into labels 1 and 0 (ignoring the neutral sentiments). 1 and 0 represents negative and positive sentiment tweets respectively. The resulting judging dataset consists of about 23k records and 3 columns which is further considered to train the model. The columns include the serial number of the tweet, the content of the tweets and the label of the sentiment of the tweet under id, tweet and Sentiment respectively.

3. **Exploratory data analysis and visualization:**

After Sentiment Analysis, the target labels are generated, and the raw dataset is decomposed into judging dataset which is about 23k records and 3 columns. The label column consists of two numerical categories, 1 and 0, where 1 indicates negative sentiments and 0 indicates positive sentiments. Also, neither duplicate tweets nor missing values are present in the dataset.

However, the data is not balanced. The dataset was moderately biased with 59 % Of the tweets or 13363 records containing positive sentiment labeled twitter data and 41% of the tweets or 9262 records containing negative sentiments labeled twitter data. The model will not be able to precisely learn which tweet is positive and which is negative when an imbalanced dataset is used for training or modeling. This leads to falsely categorizing positive tweet as negative and vice-versa.

For instance, Figure 1.1 shows distribution of positive (label 0) and negative (label 1) tweets in the dataset. There is a clear imbalance in the distribution of labels and hence it is required to balance it for better prediction results.

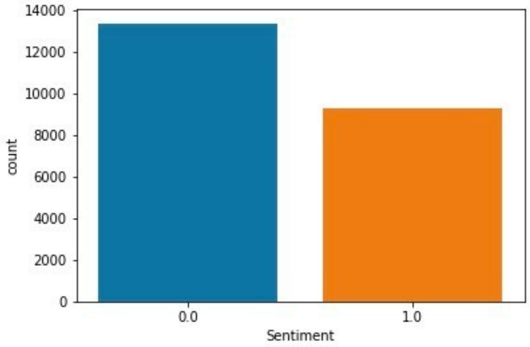


Figure 1.1: Plot of imbalanced dataset

Therefore, we perform random sampling. Here, we under sample Positive tweets as we cannot up-sample negative tweets. By taking a subset of negative and positive tweets from the actual dataset, we chose n random positive tweets, where n is the number of negative tweets and concatenate this with the subset of negative tweets. The resulting dataset is now used as the training data for the feature generation and modelling purposes. Figure 1.2 shows the balanced dataset after random sampling. Both the labels are equally distributed with each consisting of about 10k records. The code for sampling can be found in 7.2.5

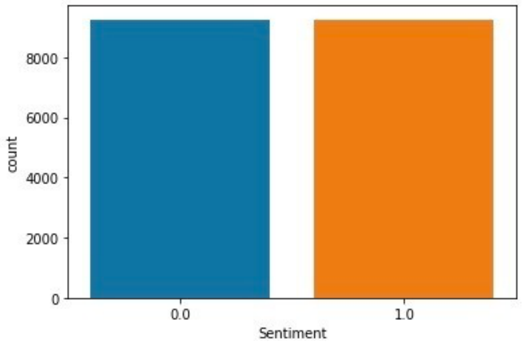


Figure 1.2: Plot of balanced dataset

Exploring and Visualizing data is an essential step in gaining insights. Hence, answering the questions regarding the data through visualization is one of the most appealing ways to solve a problem. Some of the predominant questions that needs to be answered are:

- What are the most frequently used words in the dataset?
- What are the topmost Positive and Negative words in the dataset?
- Do the hashtags throw the same sentiments as its parent tweet?

A Word Cloud is a visualization technique that can be used to visualize texts based on their frequency of use. This means that the most common words appear in large size and the less frequent words appears in smaller size. Figure 1.3 shows a word cloud with the most common words in the entire dataset. The code for word cloud can be found at 7.2.3

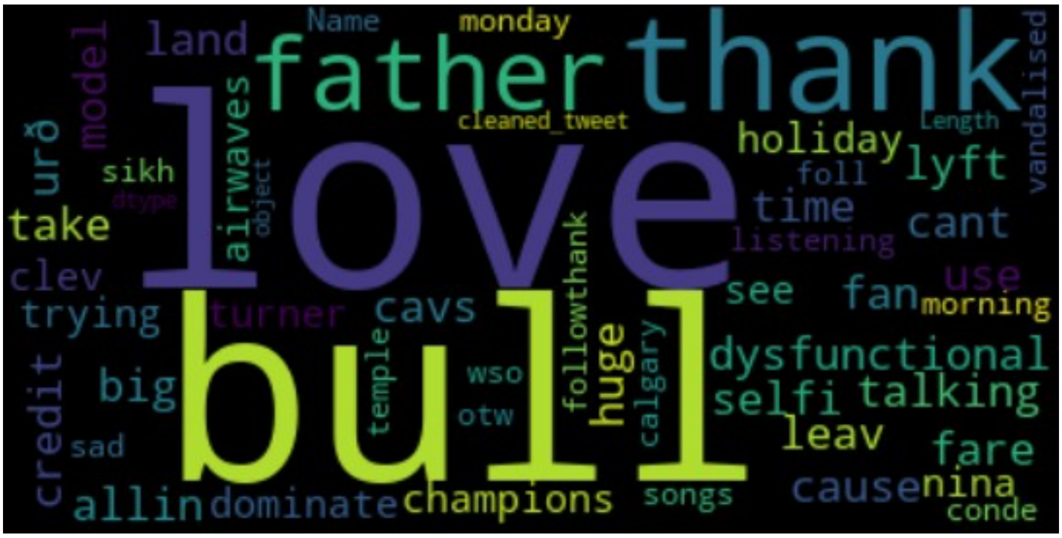


Figure 1.3: Top Words in the Twitter Hate Speech dataset

The presence of top words like love, bull, dominate, thank, dysfunctional, positive, etc. in the word cloud shows that the dataset contains a mixture of positive and negative words. In order to get a more detailed view, we plot the top positive and negative words in the dataset separately.

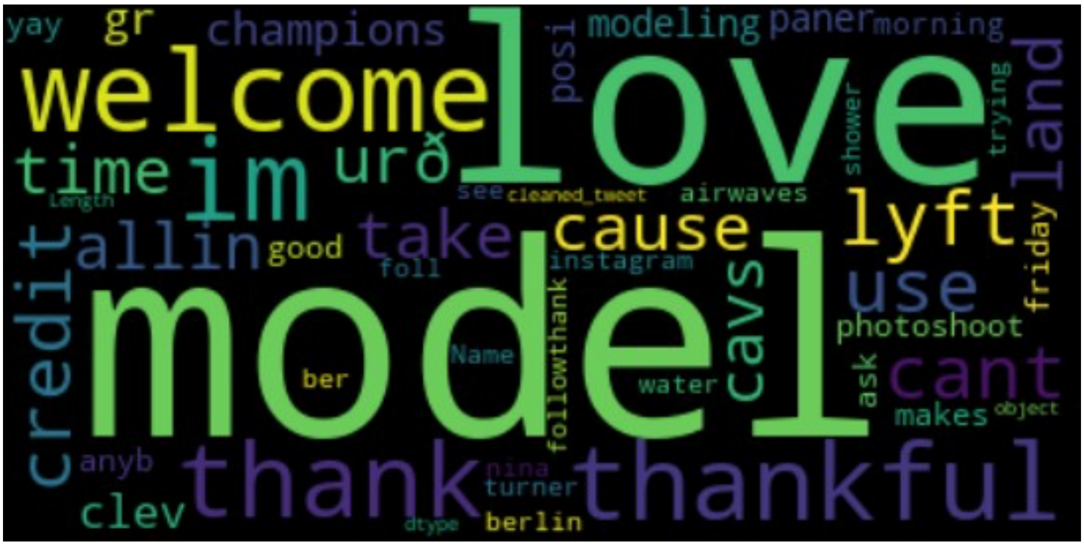


Figure 1.4: Top Positive Words in the Twitter Hate Speech dataset

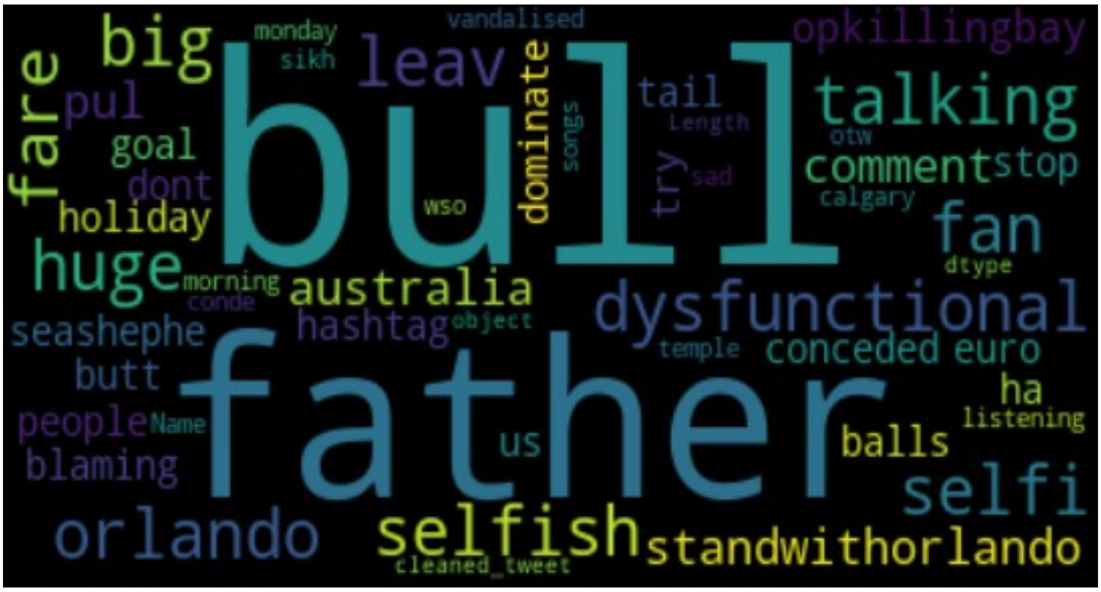


Figure 1.5: Top Negative Words in the Twitter Hate Speech dataset

The above Word Clouds are created to visualize the most frequently occurring words for two emotional labels in the tweets: Positive and Negative Sentiment tweets. In Figure 1.4, we can see that words like, love, positive, thankful, etc. shows the positive nature of the tweets and in Figure 1.5, words like bull, dominate, selfish, etc. shows the negative side of the tweets. In Figure 1.5, it can also be seen those words like father, Orlando, holiday, Australia is used in negative tweets, may be depicting some bad experience the user might have had.

The hashtag in twitter has the same meaning as the words in the tweet and hence represents the ongoing trend in social media. Therefore, checking whether these hashtags add any value to the sentiment analysis is crucial. For instance, consider the following tweet:

i am thankful for sunshine. #thankful #positive

This is a positive tweet, and the hashtags convey the same meaning.

Also, consider another example,

a transformer blew in my neighborhood. stuck w/o power, hopefully not for too long. i don't food in the fridge w/o coldness. #Blackout

This is a negative sentiment tweet where #blackout emphasize a bad experience of the user. Therefore, visualizing the top hash words is very important to understand tweet sentiments.

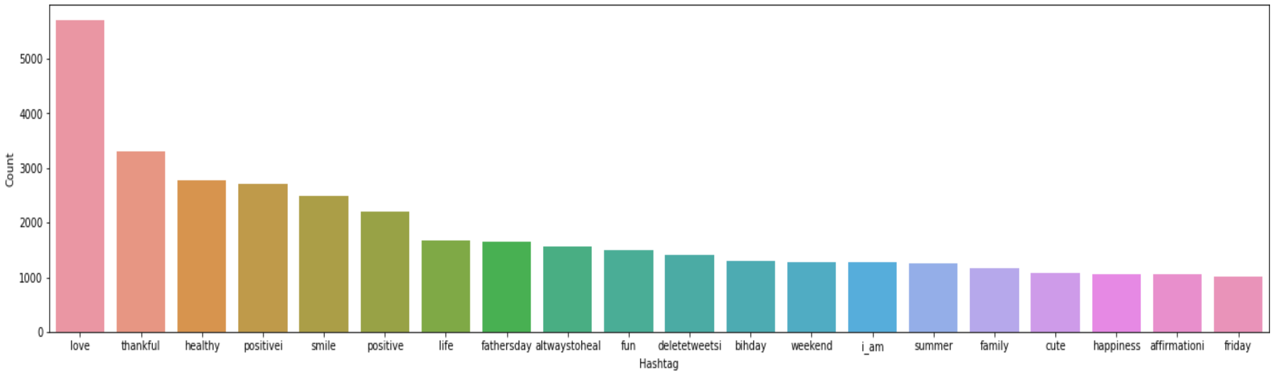


Figure 1.6: Top Positive Hashtags in the Twitter Hate Speech dataset

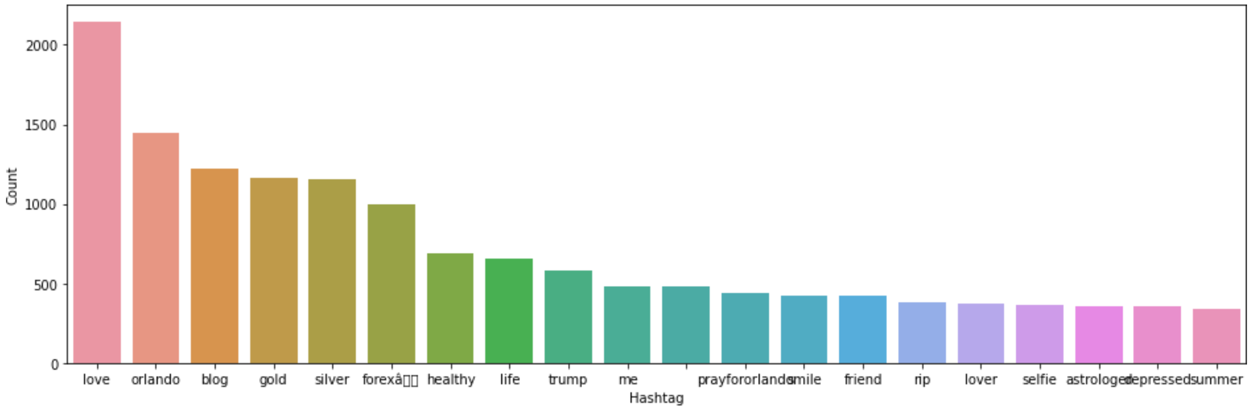


Figure 1.7: Top Negative Hashtags in the Twitter Hate Speech dataset

As we can see, in Figure 1.7, the negative hashtags have words like love, Orlando, gold, etc., most probably referencing to a bitter experience that the user might have had when he was expecting positive outcome.

The figure 1.6 shows the hashtags present in the positive sentiment labeled Tweets, where words like love, thankful, positive, etc. clearly depicts it is part of a non-negative tweet. Hence, hashtags can also be used to understand the sentiments of a twitter tweet.

4. Approach:

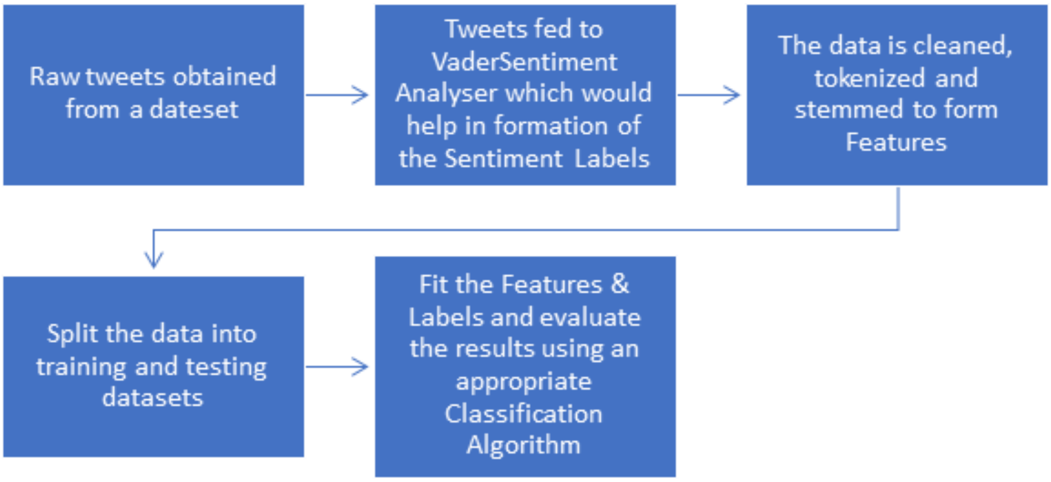

Raw tweets obtained from a dateset
Tweets fed to VaderSentiment Analyser which would help in formation of the Sentiment Labels
The data is cleaned, tokenized and stemmed to form Features
Split the data into training and testing datasets
Fit the Features & Labels and evaluate the results using an appropriate Classification Algorithm

Figure 1.8: Workflow of the approach

4.1 Creation of Labels

From the above Figure 1.8 represents the workflow of our Approach. The first step after we obtain a set of raw tweets is that we need to assign them with labels based on the overall sentiment associated with that tweet. For this purpose, we utilize Vader Sentiment Analyzer. It would help us in measuring the intensity or magnitude of a sentiment to which the tweet is specifically gravitated to. To illustrate this, let us assume that for a specific tweet, we have obtained values for the overall positivity, negativity and neutrality as 0.5, 0.2 and 0.3 respectively. The scoring mechanism works in such a way that the summation of all these values would be equal to 1. The sentiment analyzer would only indicate the intensity of sentiments by calculating these scores.

Upon this, we devise a logic to form labels based on these scores. Before we define a simple conditional logic for formulating this, we need to calculate the lower threshold values or cutoff points for determination of label values. For this, we calculate mean values of the overall sentiment scores. For instance, the mean value obtained for positive scores is 0.1322. Therefore, we build a conditional logic such that if the positive sentiment score is greater than or equal to 0.1322, then it would be assigned with label 0.

Likewise, we perform the same operation to form label 1. We obtain certain values which do not satisfy any of the above conditions. Essentially, they would be having a very strong neutral score. We drop such values from our dataset as the Vader Model itself has an ambiguity on judging the exact sentiment of that tweet as either positive or negative.

5. Analysis:

5.1. Data Cleaning

Data Cleaning is the first step of Analysis. It is imperative that we have a clean and consistent. The data is first verified for duplicates. We found no duplicate data points in the dataset. Then we check data for missing values. There were no missing values found. In order to have a uniform format of data, the data is converted into lowercase and ASCII characters. The characters that do not add value to the data is eliminated. Data cleaning code can be found in 7.2.2. The model is as good as the data that is fed into it. The steps followed in data cleaning are as follows in Figure 1.9:

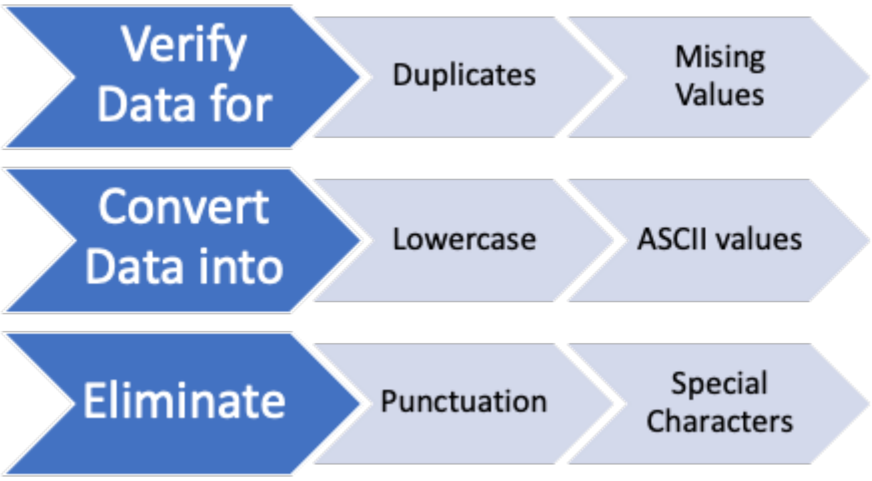


Figure 1.9: Data Cleaning Steps

5.2. Data Preprocessing & Feature Engineering

5.2.1. Data Preprocessing:

We now have obtained a dataset with the label column as a dependent variable and the tweets column as an independent variable. Our methodology to obtain processed features starts from cleaning up the tweets by removing special characters and converting the texts into lower case. We preprocess this data through three critical components of data manipulation pipeline, namely, Tokenization, Stemming and Stop Words removal as shown in the below Figure 1.10. The code for tokenization and stemming and removal of stop words can be found in 7.2.5

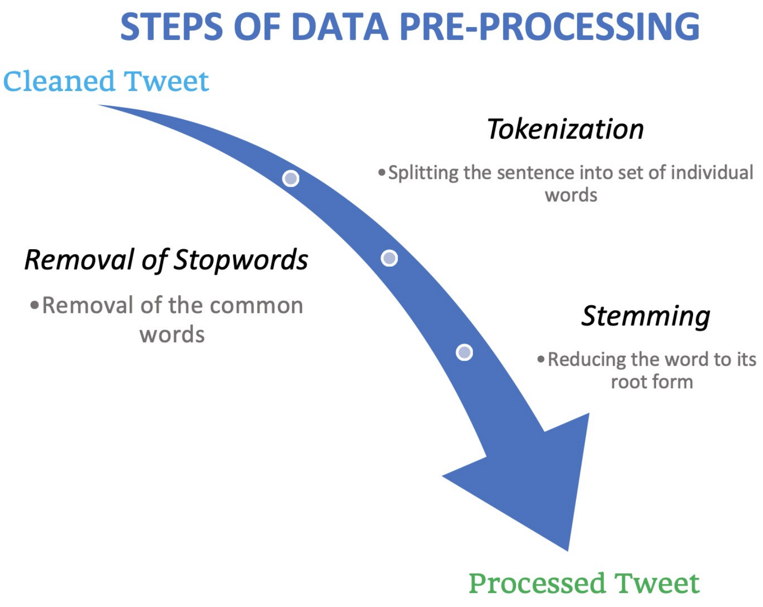

STEPS OF DATA PRE-PROCESSING
Cleaned Tweet
Tokenization
•Splitting the sentence into set of individual words
Removal of Stopwords
•Removal of the common words
Stemming
•Reducing the word to its root form
Processed Tweet

Fig: 1.10 Steps of Data Pre-processing

5.2.1.1. Tokenization

We have performed the tokenization process on the cleaned tweet data using tokenizer function as a first step for preparing our data to train the model.

For example, let us consider the following cleaned tweet,

INPUT

Cleaned_Tweet: when a father is dysfunctional and is so selfish he drags his kids into his dysfunction run

OUTPUT

Tokenized Words: [when, a, father, is, dysfunctional, and is, so, selfish, he, drags, his, kids, into, his, dysfunction, run]

5.2.1.2.Removal of Stop words

The tokenized words are further processed to remove the Stopwords and common words so that we can have on the important information from the tweet to train our model

INPUT

Tokenized_Words: [when, a, father, is, dysfunctional, and is, so, selfish, he, drags, his, kids, into, his, dysfunction, run]

OUTPUT

Meaningful_Words: [father, dysfunctional, selfish, drags, kids, dysfunction, run]

5.2.1.3 Stemming

In stemming we used Lancaster stemmer for taking the stem (root word) of the meaningful words we have in order to reducing different forms of a word to a core root word for example, the words "help" and "helped" are just different tenses of the same verb so it would be cut short to its root word help.

INPUT

Meaningful_Words: [user, father, dysfunctional, selfish, drags, kids, dysfunction, run]

OUTPUT

Stemmed_Words: [fath, dysfunct, self, drag, kid, dysfunct, run]

As a final step of preprocessing, we have combined the stemming step output to a singlesentence as a processed tweet.

Processed: fath dysfunct self drag kid dysfunct run

5.2.2. Feature Engineering:

Further, the processed tweet is used for feature engineering where the tweet texts are converted into numeric forms using the Term frequency and Inverse Document Frequency (TF-IDF) vectorizer which indicates how relevant a term is in a given document with corresponding frequency values of the words. Feature Engineering code can be found in 7.2.6

TFIDF vectorizer algorithm focusses on TF frequency of a word with respect to a document(row) and the whole corpus. Likewise, the algorithm calculates frequency of a specific word and frequency of all the documents(rows) where that word occurs.

Eventually, the texts are mapped into indices forming vectors. So as seen in the above steps, if we look at the tfidf features of the example tweet the output of tfidf vectorizer would be,

INPUT: *fath dysfunct self drag kid dysfunct run*

	fath	**dysfunct**	**self**	**drag**	**kid**	**dysfunct**	**run**	**hate**
0	0.190438	0.790007	0.339848	0.321566	0.229332	0.790007	0.246934	0.0

From, above output we can see how the tfidf vectorizer is assigning frequency values of the words present in the tweet compared with the whole corpus. The hate word frequency value is zero as the word hate is not present in the tweet, we are vectorizing, similarly all the words that are not present in the tweet would be assigned a value of zero for the word. The top five features of the tweet that we have taken as example are,

feature	**tfidf**	
0	dysfunct	0.790007
1	self	0.339848
2	drag	0.321566
3	run	0.246934
4	kid	0.229332

So, the words with highest tfidf score are in the tweet we have taken as example can be seen above.

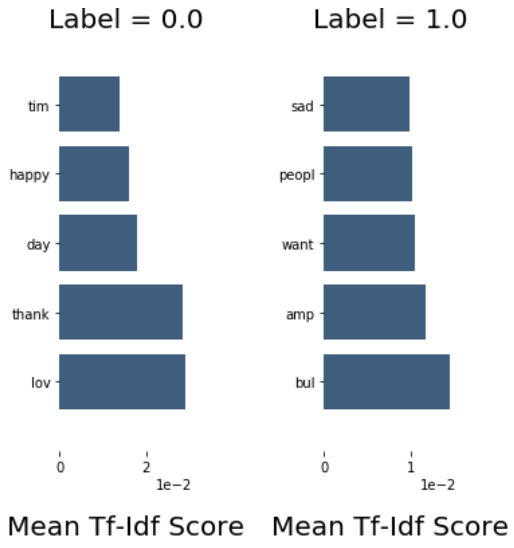


Fig: 1.11 Mean tf-idf score

Similarly, we TFIDF vectorizer calculates the frequency values for the whole processed tweets we have, and we use this as the input for our model generation.

5.3. Model Generation

Our main objective here is to build a Text Classifier which could best guess or estimate the overall sentiment of the tweets from the historical data. The vectors generated from the tweets would have certain patterns within themselves, perhaps an indication of recurring words or phrases. We expect our statistical model to capture this underlying trend within the independent words provided with their TF-IDF Scores and determine if the tweet is hate or not-hate tweets. Also, as we are aware that our dependent variable has two values, we thereby utilize the Logistic Regression algorithm for classification of tweets.

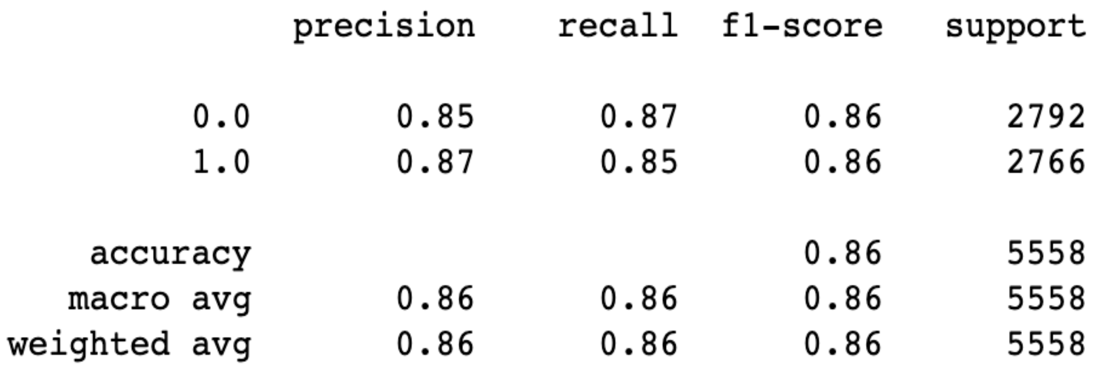
As our data is highly unbalanced, we used random sampling to balance our data. As Model greatly depends on the data fed into it.

We use sklearn package to build our logistic regression model. Logistic regression model is a statistical model, uses logistic function a S shaped curve with a function to predict the class of the independent variable. The independent variable is the bag of words we created with their TF-IDF values passed into the model as a vector. The dependent variable for our model is the Sentiment classification label we obtained by using Vader sentiment analysis.

We split our data into training and testing datasets in the ratio 70:30. The training dataset is used to train the Model. The model applies linear regression and a logistic function to determine whether a tweet is hateful or not- hate tweet. Code for model generation and building can be found at 7.2.8

5.4. Result Analysis:

The result analysis of the model built to predict the hate or no-hate tweet can be summarized in the table below. The accuracy, precision, recall and f1 score indicates that our model is a pretty good model. The model can predict whether a tweet is hateful of not hateful with 84% accuracy. Code for model Evaluation can be found at 7.2.9



```
              precision    recall  f1-score   support

         0.0       0.85      0.87      0.86      2792
         1.0       0.87      0.85      0.86      2766

    accuracy                           0.86      5558
   macro avg       0.86      0.86      0.86      5558
weighted avg       0.86      0.86      0.86      5558
```

The ROC curve for the model summarizes the trade-off between the true positive rate and false positive rate for this predictive model and the Figure 1.13 indicates that we have a pretty good and balanced model.

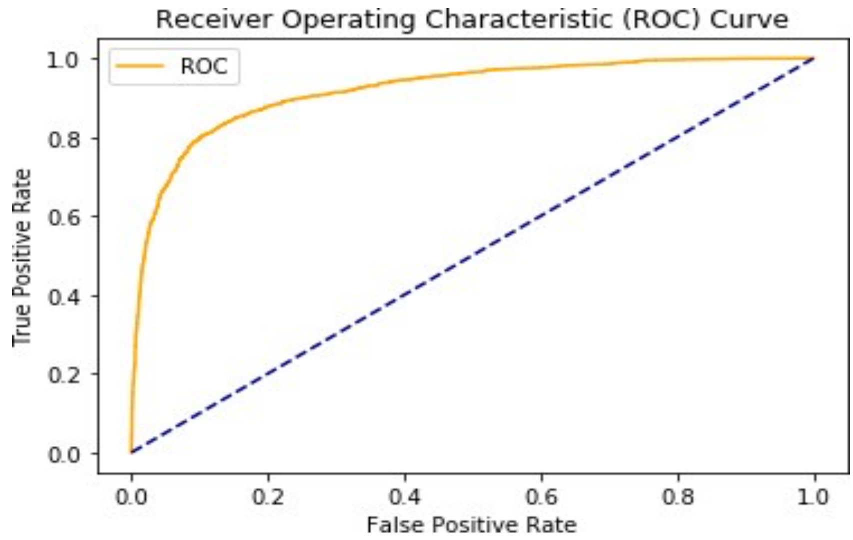


Fig: 1.13 ROC Curve

6. Summary

All in all, we have observed how different numerical metrics could be used to evaluate performance of a binary text classifier. We managed building an end-to-end pipeline for critical NLP processes, right from data extraction to model generation. To briefly summarize the proceedings, we have obtained a fairly accurate statistical model to predict the sentiment of the tweets.