NLP Project MS1

Mohamed Ashraf

May 8, 2023

1 Problem

This task's objective is to find instances of hate speech in tweets. We characterise a tweet as containing hate speech if it expresses racial or sexist views in order to make the process easier. The goal is to differentiate between tweets that include these sentiments and those that do not.

2 Data Collection

We will impelement a supervised model, so we need a dataset that could be divided into two independent sets. The data was obtained from an open source dataset that was made accessible on Kaggle.

- Test set
- Train set

3 Data Size and Structure

data is splitted into two main sets

- Test dataset (31962 rows)
- Train dataset (17197 rows) The following illustration demonstrates how data is organised into columns to assist in resolving the sentiment analysis problem.

4 Tweets Preprocessing and Cleaning

Text is an inherently unstructured type of data, characterized by a variety of noise types that render it challenging to analyze without prior processing. Text preprocessing involves cleaning and standardizing the text by removing noise and preparing it for analysis. This task will be split into two sections:

- Data Inspection
- Data Cleaning

id	label	tweet
14	1	@user #cnn calls #michigan middle school 'build the wall' chant '' #tcot
15	1	no comment! in #australia #opkillingbay #seashepherd #helpcovedolphins #thecove #helpcovedolphins
18	1	retweet if you agree!
24	1	@user @user lumpy says i am a . prove it lumpy.
35	1	it's unbelievable that in the 21st century we'd need something like this. again. #neverump #xenophobia
57	1	@user lets fight against #love #peace

Figure 1: Data sample.

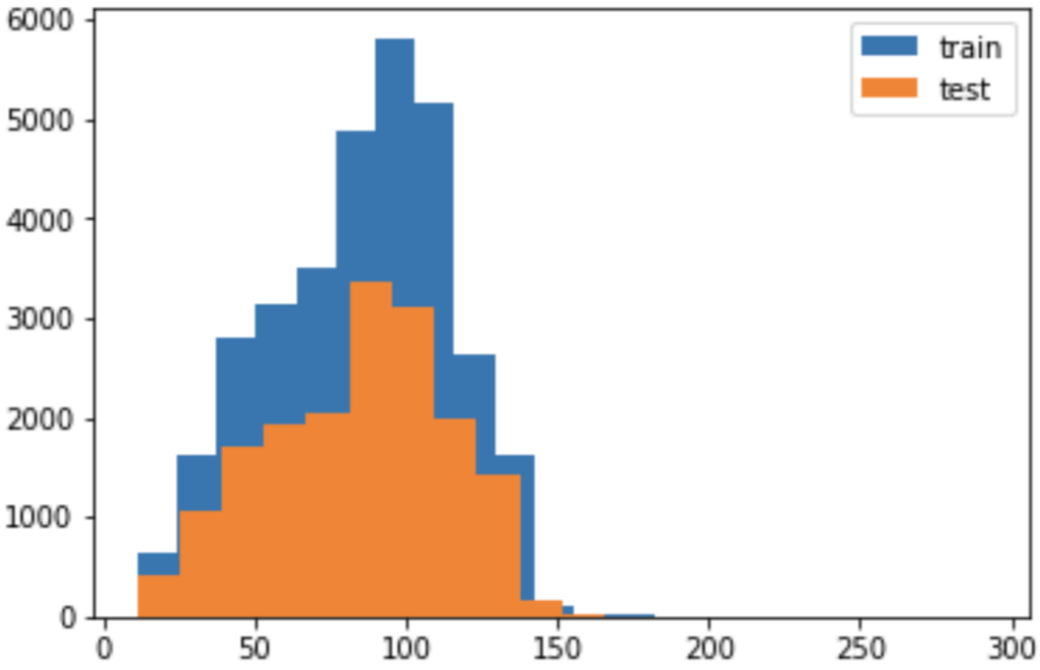

Figure 2: length of the tweets.

4.1 Data Visulaization

This graph illustrates the distribution of tweet lengths in terms of words for the train and test sets of data.

4.2 Data Cleaning

In this step we will do some data preprocessing in order to obtain data that can be used to extract features to help in modeling our problem such as

- Remove Twitter handles
- Remove punctuations, numbers and special characters
- Remove small words such as (the,and,we)
- Tokenization of Sentences
- Normalize the text (Stemming)
- Remove small words such as (the,and,we)

4.3 Data Analysis

As hashtags are commonly employed as keywords in tweets on Twitter, we will generate two sets of hashtags that comprise both positive and negative emotions.

4.4 Feature extraction (tf-idf)

The TF-IDF algorithm diminishes the importance of common terms by assigning them lower weights, and elevates the significance of words that are infrequent across the entire corpus but appear frequently in a small subset of documents.

- TF = (Number of times term t appears in a document)/(Number of terms in the document)

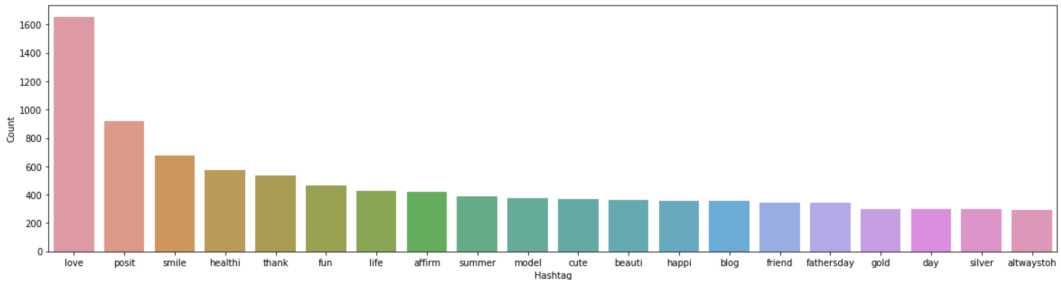

Figure 3: Good sentiments example.

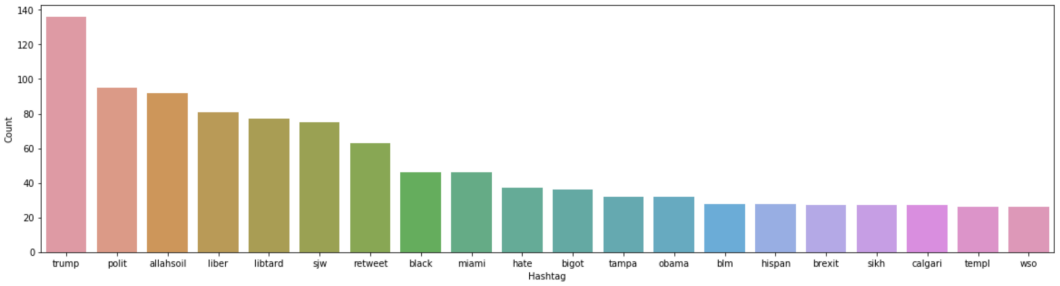

Figure 4: bad sentiments example.

- IDF = log(N/n), where, N is the number of documents and n is the number of documents a term t has appeared in.
- TF-IDF = TF*IDF

4.5 System Architecture Flow

We will be using our TF-IDF values inside the Matrix as weights assigned to each feature to be used in our Logistic regression model

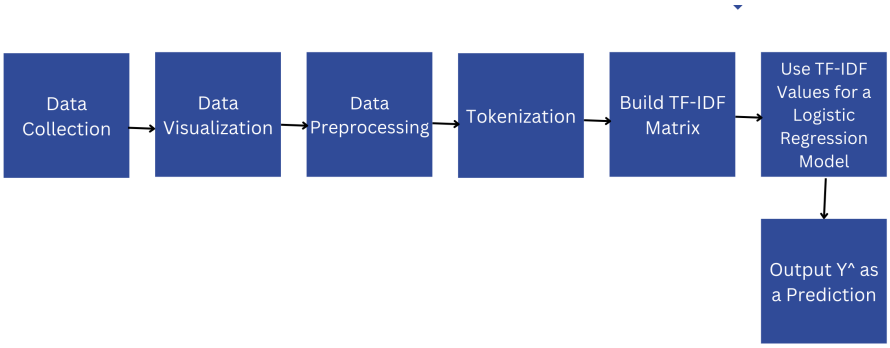

Figure 5: System Flow.