NLP PROJECT

MS1

MOHAMED ASHRAF

1. UNDERSTAND THE PROBLEM

1. THE OBJECTIVE OF THIS TASK IS TO DETECT HATE SPEECH IN TWEETS. FOR THE SAKE OF SIMPLICITY, WE SAY A TWEET CONTAINS HATE SPEECH IF IT HAS A RACIST OR SEXIST SENTIMENT ASSOCIATED WITH IT. SO, THE TASK IS TO CLASSIFY RACIST OR SEXIST TWEETS FROM OTHER TWEETS.

DATA COLLECTION

DATA VISUALIZATION (EXPLORATION)

DATA CLEANING

FEATURE EXTRACTION

1. DATA COLLECTION

THE DATA USED WAS COLLECTED FROM AN OPENSOURCE DATASET AVAILABLE ON KAGGLE WE WILL IMPLEMENT A SUPERVISED MODEL SO WE NEEDED A DATASET THAT WE CAN BE SPLITTED INTO TWO DIFFERENT SETS

1. TEST DATASET
2. TRAIN DATASET

DATA SIZE AND STRUCTURE

DATA IS SPLITTED INTO TWO MAIN SETS

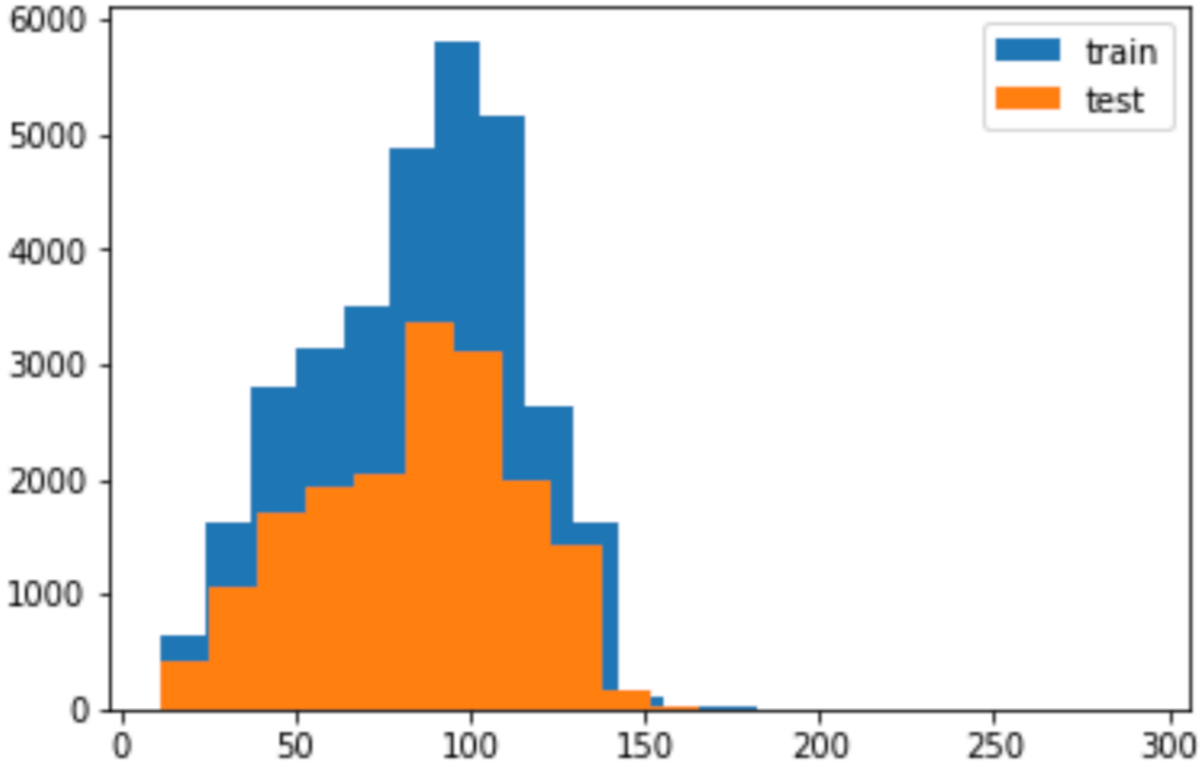
1. TEST DATASET (31962 ROWS)
2. TRAIN DATASET (17197 ROWS)

The Data will be in the following format

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

DATA VISULAIZATION

this graph shows the distribution of length of the tweets, in terms of words, in both train and test data.

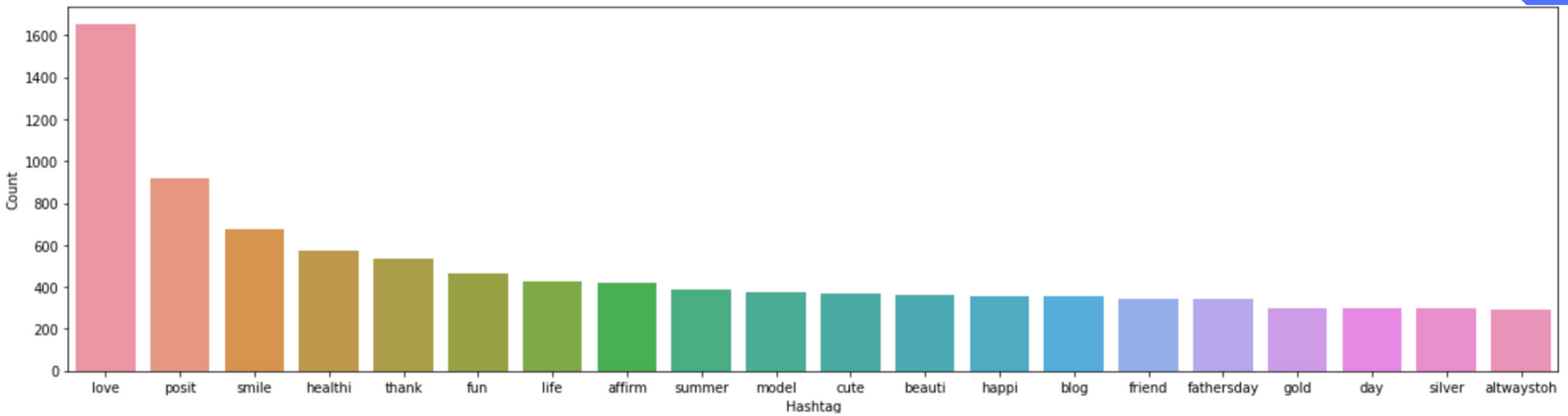


DATA CLEANING

1. Remove the Twitter handles
2. Remove punctuations, numbers and special characters
3. Remove small words such as (the,and,we)
4. Tokenization of Sentences
5. Normalize the text (Stemming)

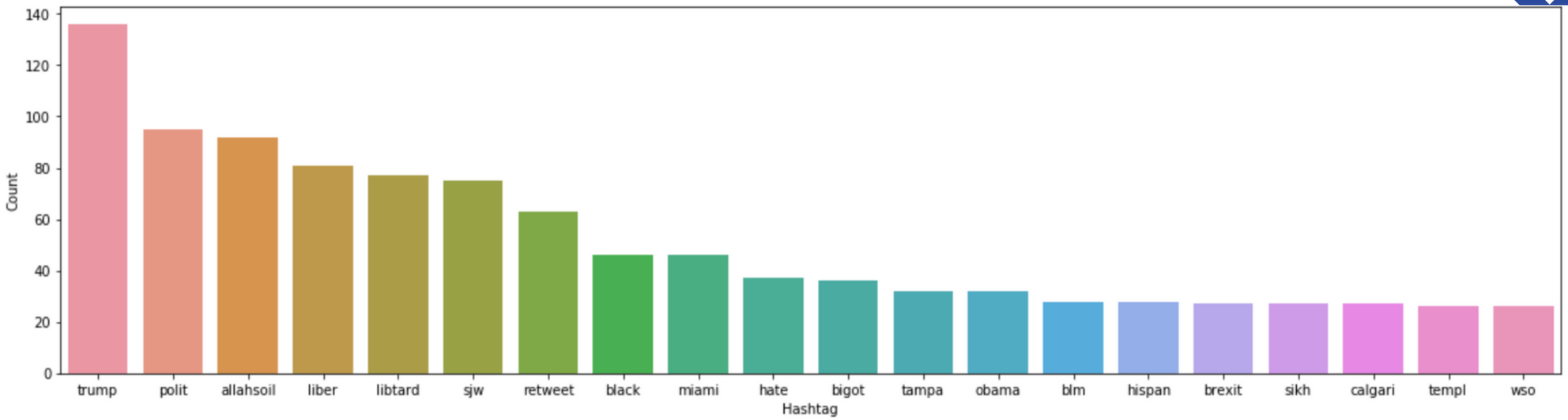
DATA ANALYSIS

Hashtags in Twitter are commonly used as keywords to the tweets posted so we will create two sets that contains both the Positive sentiment Hashtags and the negative sentiment ones



Positive Sentiment Hashtags Count

DATA ANALYSIS



Negative Sentiment Hashtags Count

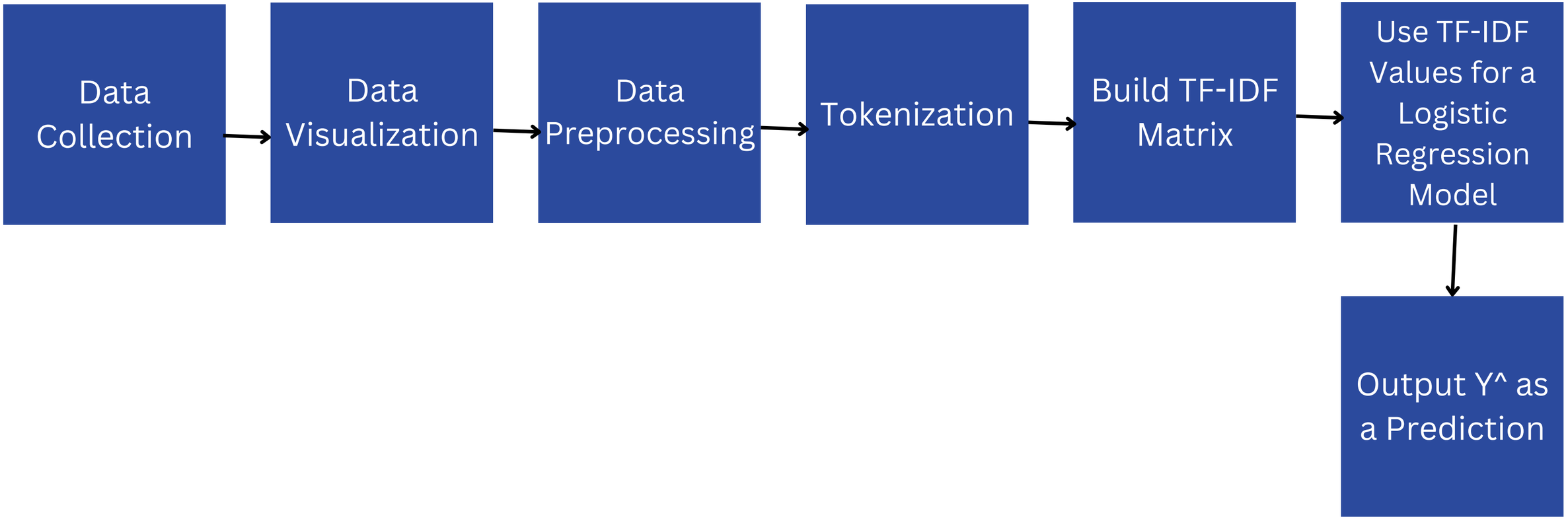
FEATURE EXTRACTION (TF-IDF)

TF-IDF works by penalising the common words by assigning them lower weights while giving importance to words which are rare in the entire corpus but appear in good numbers in few documents.

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

Logistic regression model will be used to predict a binary outcome given a set of independent variables

SYSTEM ARCHITECTURE FLOW



THANK YOU

Mohamed Ashraf
46-0831