NLP Assignment-1 REPORT

Team members:- **Harjeet Singh Yadav** - 2020561, **Aaditya Gupta** - 2020552

Section I:

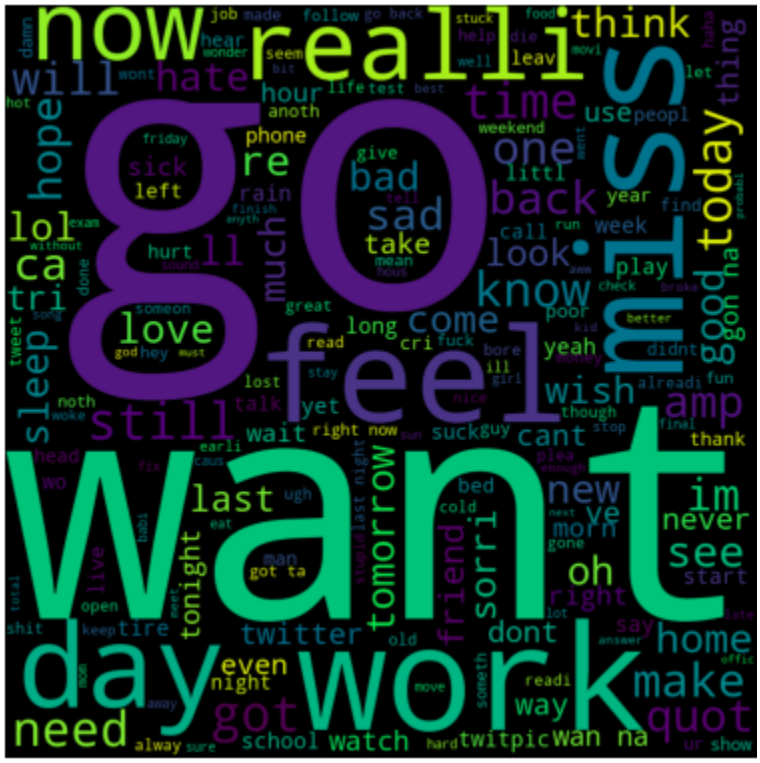
- We have used the regex re library of python for finding the pattern.
- We have assumed that any word followed by a full stop, comma,? Mark the ending of a sentence and we have split the tweets into sentences based on this observation.
- \w+[^\w\s]+ we have used this regex to count the number of tokens because this also counts punctuation as separate tokens. eg Hello, -> 'Hello', ',' 2 tokens
- Lowercase the tweets in part c and have used lowercase text for further processing.
- `@\w{1,15}\s` used this to match usernames as they can only be 15 characters and can contain only alphanumeric and underscore characters.
- We have assumed that URLs will either start with https?: or www.
- `\b[aeiouAEIOU]\w*` utilised /b to mark word boundary in the B part.

Section II:

- We have utilised the NLTK library for text preprocessing
- We have used various libraries in NLTK like `word_tokenize for tokenization, SnowballStemmer for stemming WordNetLemmatizer For lemmatization and stopwords for nltk.corpus for getting stopwords`
- substitution of re (regEx) library method has been used to remove the stopwords, URLs and HTML tags from the tweets

Section III:

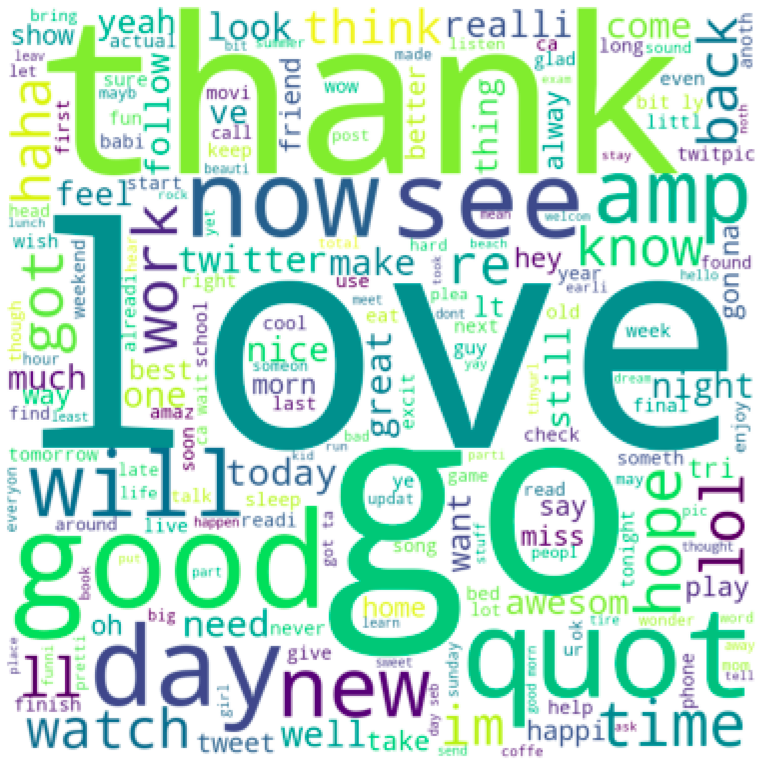
- Used the word cloud for the text visualization
- Below is the text visualization for class 0 i.e negative tweets



Observation:

Most frequently used words in negative tweets are go, feel, want, work, bad, sad, take, hope, leave, lost, cant, sorri etc. most of these words actually reflect negative sentiment.

- Below is the text visualization for class 1 i.e positive tweets



Observation:
Most frequently used words in class 1 i.e positive tweets are thank, love, good, go, good, want, lol, new, look, haha, happi, help, time, etc these words actually carry a positive sentiment and feeling.

Section IV:

- We have used VADER sentiment analysis to generate labels for the given tweets. The VADER analyser gives the probability of a tweet being positive, negative and neutral.
- Finally, to predict the final label, we use the value of compound. When compound is >=0.05, we give the tweet a label of 1 i.e. positive sentiment. When compound is <= -0.05, we give the tweet a label of 0 i.e. negative sentiment.
- To predict the accuracy of the model, we calculate the true positives predicted by VADER analyzer and divide it by the total sample.

Observations:
The accuracy of predicting positive tweets after pre-processing is 64.6% and before is 62.3%
The accuracy of predicting negative tweets after pre-processing is 43.6% and before is 42%.

Contribution:-
Aaditya Gupta:- Section I and Section IV
Harjeet Singh Yadav:- Section II and Section III