Submitted by : Kapuriya Janakkumar
Sureshbhai (MT22032)
Shubham Agrawal
(MT22124)

Assumptions:

1. Username starts with ‘@’
2. Sentences end with either “.”, “!” or “?” followed by one or more white space
3. Punctuations are considered as separate tokens
4. Words starting with vowels are at least of length 2

Methodology:

1. Regex used for sentence splitting: "(?<![A-Z][a-z]\.)(?<![A-Z][a-z][a-z]\.)(?<=\.|\!|\?)\s+"

   Explanation: It splits sentences on ‘\s’ provided the look behind is either full stop, comma or exclamation. Additionally, it can also handle titles like “Mr.”, ”Mrs.”, “Dr.”, ”Jr.” etc, i.e. it won't split the sentences on periods contained within titles.

2. Tokenization using regex:

   Firstly, we found all punctuations by iterating through tweets using regex except period.
   As period(“.”) can be a part of emails, ip addresses etc, every period may not be a separate token and is handled separately.
   We found all the full stops through the regex "\w+\.\s". These are separate tokens, which are added to the list of tokens.
   After removing all the punctuations, we now split every tweet using “\s+” which gives us words which are added to the list of tokens.

3. Regex for vowels: “"(?<=^|(?<=[\s+]))[aeiouAEIOU][^\s+](?=[\s+]|$)"”

   Explanation: If look-behind is “start of string” or whitespace(one or more), followed by a vowel, then the word starts with a vowel and is either at the start of the string or in the middle. It must be followed by at least one non white space character followed by spaces or “end of string”.

4. Regex for username: “"@[a-zA-Z0-9_]+"”
   Explanation: starts with “@” followed by any number of alpha-numneric characters (at least one)

5. Regex for URL: "https?://[a-zA-Z0-9_\?=\@\/#=.~-]+"
   Explanation:Url starts with http or https followed by :// and any alphanumeric character or special symbols like (~, ?, @, #, =, . ~) Which are non white space characters.

Analysis:

1. The count of URLs in the positive class is almost double that of the negative class. This may be due to people recommending a particular website/e-resource and sharing their URL.

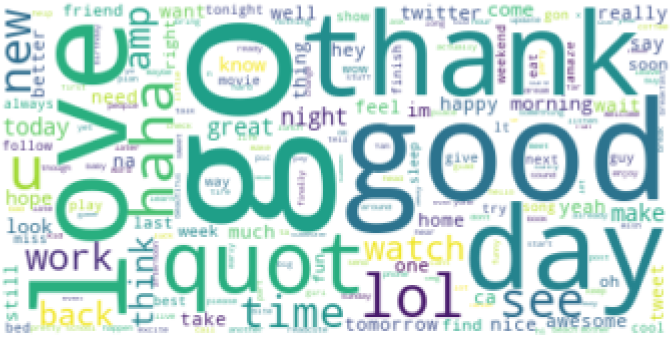
2. The number of tweets on Sundays are very high, implying people spend more time on twitter on Sundays when they probably have a holiday.

3. Vader sentiment analysis:

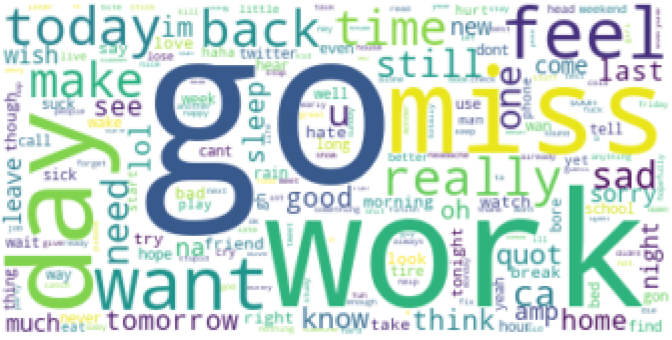
   accuracy of raw data = 65.2, accuracy of preprocessed data = 63.6. The drop in accuracy may be due to loss in information due to pre-processing. Removal of punctuations, case folding may dampen the sentiment. Ex. “Amazing work!!!!!”(raw-text) Has a stronger sentiment vs “amazing work”(pre-processed text)

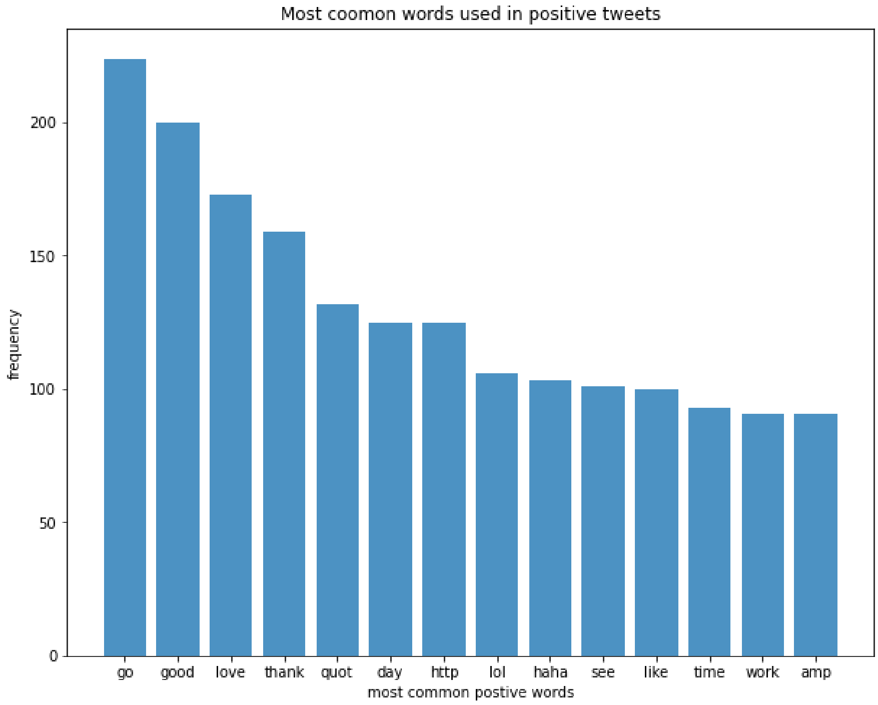
4. Word Clouds:

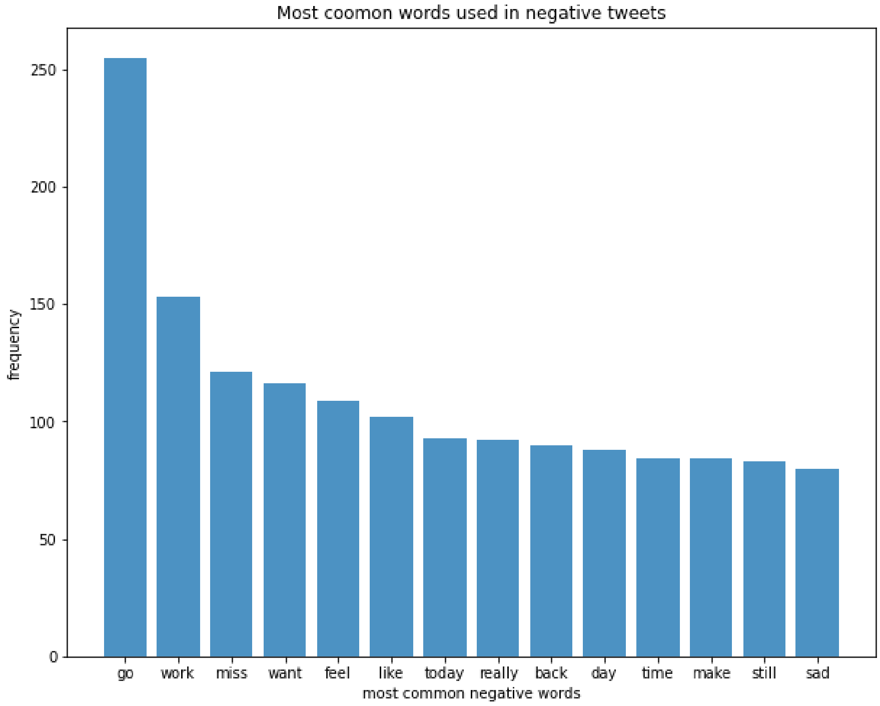
   Positive Word Cloud



   Negative Word Cloud







Some words occur frequently in both positive and negative classes (go, work, back, and time). These words affect the sentiment based on the context.

People generally have negative sentiments w.r.t “work” which may point to a general sense of dissatisfaction amongst the working class. Frequent occurrence of “want” in tweets with negative sentiment may depict dissatisfaction and discontent among people

Helper_Functions:

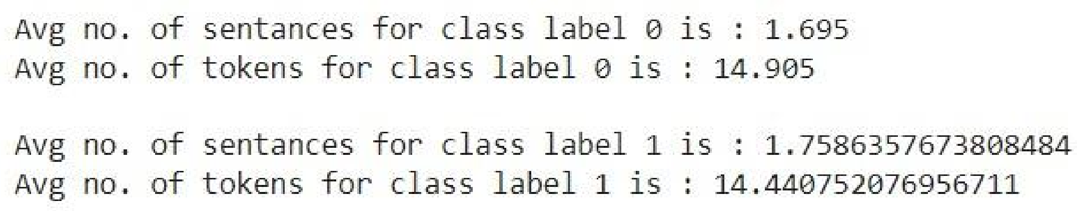
def accuracy(size, trueVSguess) - returns the accuracy of Vader sentiment analyzer on the given dataset. Input: total number of tweets and list of tuples containing true and predicted labels.

Preprocessing_Steps:

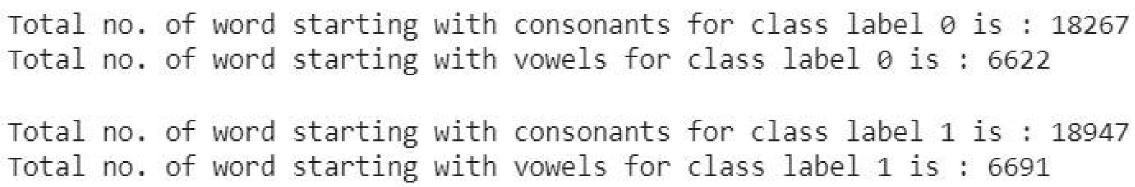
```
A. White space removal
B. Case Folding
C. Tokenization
D. Stop words removal
E. Punctuation removal
F. Url removal
G. Spelling check
H. Lemmatization
```

Contributions: We equally divided the work such that both of us get to work on everything. Both of us were equally involved in this work.

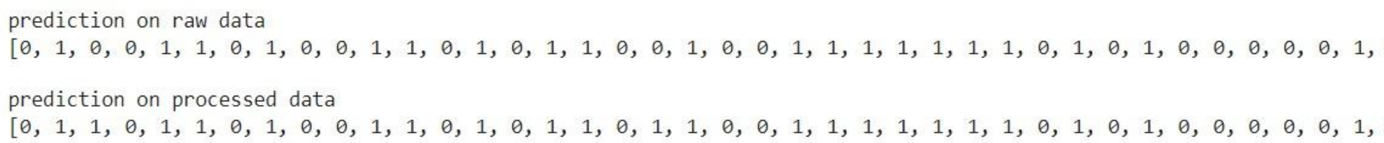
Outputs:



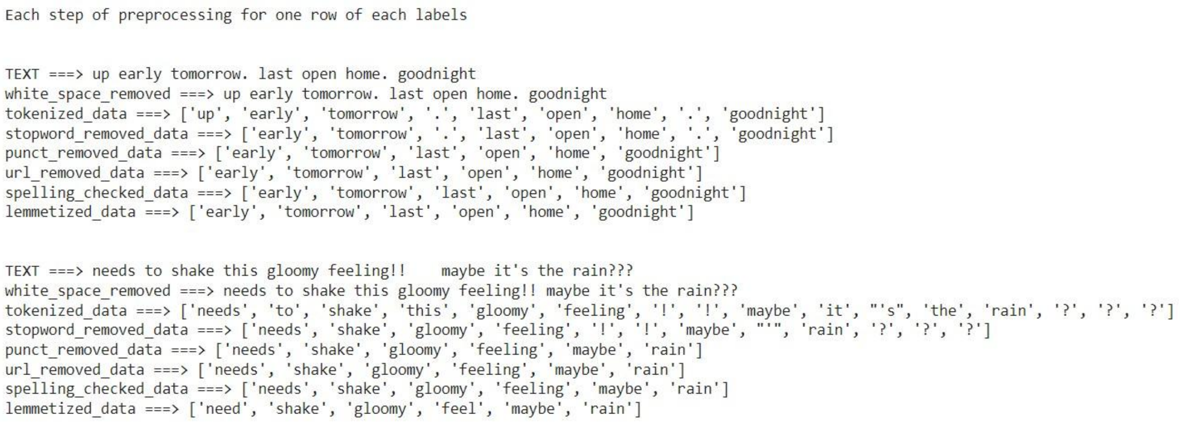
```
Avg no. of sentances for class label 0 is : 1.695
Avg no. of tokens for class label 0 is : 14.905

Avg no. of sentances for class label 1 is : 1.7586357673808484
Avg no. of tokens for class label 1 is : 14.440752076956711
```



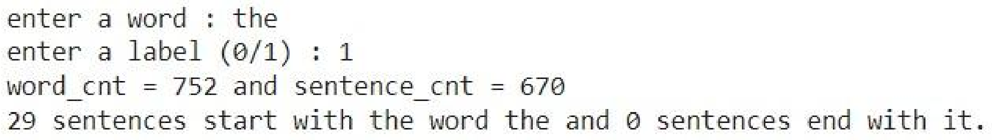
```
Total no. of word starting with consonants for class label 0 is : 18267
Total no. of word starting with vowels for class label 0 is : 6622

Total no. of word starting with consonants for class label 1 is : 18947
Total no. of word starting with vowels for class label 1 is : 6691
```



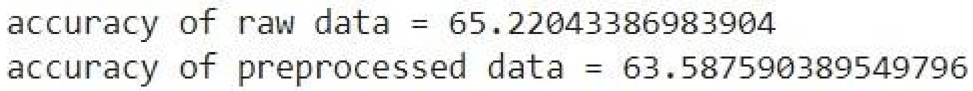
```
prediction on raw data
[0, 1, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1,

prediction on processed data
[0, 1, 1, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1,
```



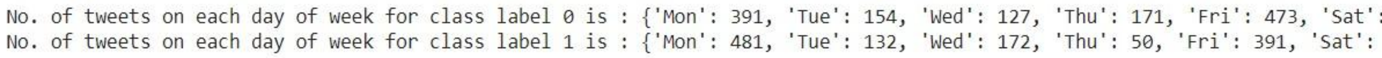
```
Each step of preprocessing for one row of each labels


TEXT ===> up early tomorrow. last open home. goodnight
white_space_removed ===> up early tomorrow. last open home. goodnight
tokenized_data ===> ['up', 'early', 'tomorrow', '.', 'last', 'open', 'home', '.', 'goodnight']
stopword_removed_data ===> ['early', 'tomorrow', '.', 'last', 'open', 'home', '.', 'goodnight']
punct_removed_data ===> ['early', 'tomorrow', 'last', 'open', 'home', 'goodnight']
url_removed_data ===> ['early', 'tomorrow', 'last', 'open', 'home', 'goodnight']
spelling_checked_data ===> ['early', 'tomorrow', 'last', 'open', 'home', 'goodnight']
lemmetized_data ===> ['early', 'tomorrow', 'last', 'open', 'home', 'goodnight']


TEXT ===> needs to shake this gloomy feeling!!    maybe it's the rain???
white_space_removed ===> needs to shake this gloomy feeling!! maybe it's the rain???
tokenized_data ===> ['needs', 'to', 'shake', 'this', 'gloomy', 'feeling', '!', '!', 'maybe', 'it', "'s", 'the', 'rain', '?', '?', '?']
stopword_removed_data ===> ['needs', 'shake', 'gloomy', 'feeling', '!', '!', 'maybe', "'", 'rain', '?', '?', '?']
punct_removed_data ===> ['needs', 'shake', 'gloomy', 'feeling', 'maybe', 'rain']
url_removed_data ===> ['needs', 'shake', 'gloomy', 'feeling', 'maybe', 'rain']
spelling_checked_data ===> ['needs', 'shake', 'gloomy', 'feeling', 'maybe', 'rain']
lemmetized_data ===> ['need', 'shake', 'gloomy', 'feel', 'maybe', 'rain']
```



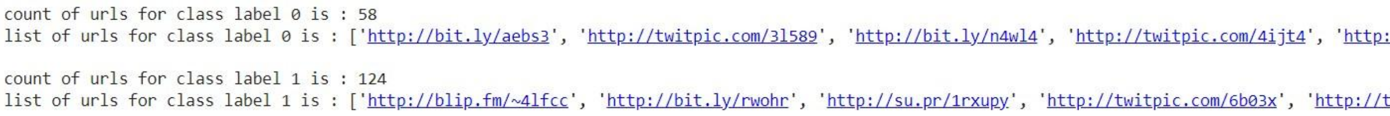
```
enter a word : the
enter a label (0/1) : 1
word_cnt = 752 and sentence_cnt = 670
29 sentences start with the word the and 0 sentences end with it.
```



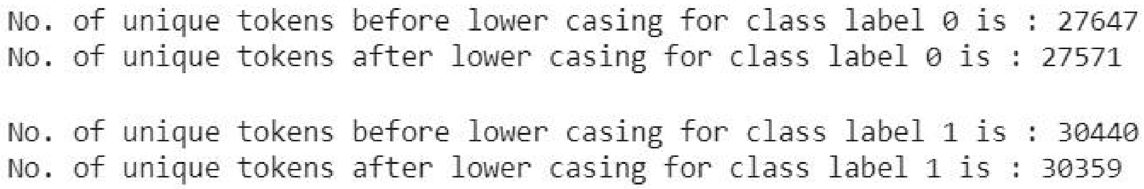
```
accuracy of raw data = 65.22043386983904
accuracy of preprocessed data = 63.587590389549796
```



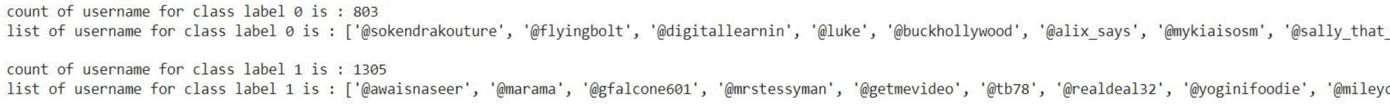
```
No. of tweets on each day of week for class label 0 is : {'Mon': 391, 'Tue': 154, 'Wed': 127, 'Thu': 171, 'Fri': 473, 'Sat':
No. of tweets on each day of week for class label 1 is : {'Mon': 481, 'Tue': 132, 'Wed': 172, 'Thu': 50, 'Fri': 391, 'Sat':
```



```
count of urls for class label 0 is : 58
list of urls for class label 0 is : ['http://bit.ly/aebs3', 'http://twitpic.com/3l589', 'http://bit.ly/n4wl4', 'http://twitpic.com/4ijt4', 'http:

count of urls for class label 1 is : 124
list of urls for class label 1 is : ['http://blip.fm/~4lfcc', 'http://bit.ly/rwohr', 'http://su.pr/1rxupy', 'http://twitpic.com/6b03x', 'http://t
```



```
No. of unique tokens before lower casing for class label 0 is : 27647
No. of unique tokens after lower casing for class label 0 is : 27571

No. of unique tokens before lower casing for class label 1 is : 30440
No. of unique tokens after lower casing for class label 1 is : 30359
```



```
count of username for class label 0 is : 803
list of username for class label 0 is : ['@sokendrakouture', '@flyingbolt', '@digitallearnin', '@luke', '@buckhollywood', '@alix_says', '@mykiaisosm', '@sally_that_

count of username for class label 1 is : 1305
list of username for class label 1 is : ['@awaisnaseer', '@marama', '@gfalcone601', '@mrstessyman', '@getmevideo', '@tb78', '@realdeal32', '@yoginifoodie', '@miley
```