**A gist**

1. As a part of this assignment, we have collected raw twitter data, tweets and clustered people into groups.

2. We were able to analyze TextBlob, Afinn dict for sentiment classification (Sentiment analysis of a given tweet).

3. Furthermore, we had data which was labelled through AFINN dict, hence we made use of the data for building the prediction model, and then splitting them into 70% of Training data and 30% testing data.

6. Vectorized the training and testing data, fitted the Logistic regression on this training data.

7. Predicted the Sentiment analysis for the testing data.

8. Found the accuracies of our model prediction.

**Description in detail:**

**Collect the raw data:**

|  |
| --- |
|  |

This module will collect the raw data from twitter and creates output pickle files for further processing. As a part of this project, the file does the following:

a.) Establish twitter connection through tokens, and create a tweepy API for api calls.

b.) Reads the input file with movie names. This assignment focusses on getting tweets and response object based on a single word – Amazon (with respect to current financial trends).

c.) Search API query is made on the word – “AMAZON” and response object is gathered.

e.) Collect the tweets from the response object and store in -----------> Tweets.pkl file for processing in **Classify.py** file (another file where sentiments are observed).

**Classifying the data – Sentiment analysis of a text or a tweet in this case.**

This file reads the collected tweets, cleans the text by removing any https reference, punctuations, stop\_words, languages other than english and sends the clean text for sentiment analysis.

a.) The collected tweets are read, and the text are pre-processed by using ***regex*** functions to eliminate https reference, punctuations etc.

b.) The cleaned text is then analyzed via ***TextBlob*** and ***Afinn*** dataset (baselined models).

c.) The ***textBlob*** is a library where we send the complete single cleaned text to the function, and it gives us the polarity of the sentence. If the polarity is > 1, then it is considered to be a positive sentiment text. If the polarity is < 1, then the tweet is negative, If the polarity is == 0, then it is a neutral tweet.

d.) The percentage of pos, neg, neutral tweets are calculated and results are written to ------------> "Classify\_results.txt" file. (Sample of the report file attached below)

e.) Another method of sentiment classification is via, ***AFINN*** dict. Here, the tweet is tokenized and each word is analyzed against the AFINN dict. If the word is present in the AFINN list of words, based on its label, we find out if the whole tweet is positive or negative or neutral.

f.) The percentage of the above process is also calculated and the results are written to the same output file --------------> "Classify\_results.txt" file.

g.) Now, having baselined, the afinn results with data and its corresponding labels, we have built a classifier model using Logistic regression.

h.) Creating a dataframe to store the positive, negative and sentiment polarity along with tweets.

i.) The data is then split into training and testing. The split ratio of training and testing is 70% and 30% respectively.

j.) Unskew the data according to the length of positive tweets retrieved from afinn\_data above. Such that training data will have more or less equal number of positive, negative, neutral length.

k.) Training data is then vectorized to get the vectors and its corresponding vocab. Initially the vocab will be none.

l.) Testing data is then vectorized to get the vectors and the vocab obtained from the above step is obtained to calculate.

m.) Use Logistic regression model and fit the training data.

***Logistic regression*** is used to describe data and to explain the relationship between one dependent binary variable and one or more independent variables.

n.) Predict the Sentiment analysis (Positive, Negative or neutral) for testing data by passing the tweets.

o.) Record the results in =--------------------> Classify\_results.txt

***Classify\_results.txt for the word – AMAZON***

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*Analyzing the tweets collected from collect file\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  
  
The Number of Tweets collected : 63  
  
\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  
 SENTIMENT ANALYSIS VIA TEXTBLOB FOR TWEETS COLLECTED   
\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  
  
Number of Positive instances found: 9  
Number of Negative instances found : 13  
Number of Neutral instances found : 41  
  
Percentage of Positive tweets: 14.285714285714286 %  
Percentage of Negative tweets: 20.634920634920636 %  
Percentage of Neutral tweets: 65.07936507936508 %  
  
Example of a Positive class : love goes on dreams come true h  
Example of a Negative class : rt what s a 63 year old man with bad hips to do when attacked by bloodthirsty aliens hide scifi contemporary 5star kindl  
Example of a Neutral class : rt ever 4 ever twitter rt amazon 5000 2 1000 10  
  
\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  
 SENTIMENT ANALYSIS VIA AFINN DATASET FOR TWEETS COLLECTED   
\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  
  
Number of Positive instances found: 9  
Number of Negative instances found : 14  
Number of Neutral instances found : 40  
  
Percentage of Positive tweets: 14.285714285714286 %  
Percentage of Negative tweets: 22.22222222222222 %  
Percentage of Neutral tweets: 63.492063492063494 %  
  
Example of a Positive class : it s great  
Example of a Negative class : rt life hack if you ever feel sad think of an animal any animal and go to amazon and search plush animal and read th  
Example of a Neutral class : rt ever 4 ever twitter rt amazon 5000 2 1000 10  
  
\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  
 CLASSIFIER BUILT ON AFINN DATA AND ITS LABELS USING LOGISTIC REGRESSION MODEL   
\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  
The tweets are split into train(70%) and testing(30%) data and the scores are caluclated as below.  
  
  
Positive tweets are high. Hence, for building classifier, the tweets in the training set are slightly oversampled to match the number of positive occurences.  
The score for trained data: 0.648649   
The score for tested data: 0.615385   
  
\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Trained data calculations below \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  
  
Total row count in trained data: 37  
  
Number of Positive Instances: 6   
Number of Negative Instances: 24   
Number of Neutral Instances: 7   
  
Example of a Positive tweet: love goes on dreams come true h   
Example of a Negative tweet: rt breaking bad 2008 cinematography by reynaldo villalobos directed by tricia brock buy the complete series via   
Example of a Neutral Tweet: rt warped passages by is on sale for 2 99 via grab it via   
  
\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Testing data calculations below \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  
  
Total row count in testing data: 26  
  
Number of Positive Instances: 0   
Number of Negative Instances: 26   
Number of Neutral Instances: 0