Twitter Sentiment Analysis of News Articles

By Gaurang Davda
College of Engineering
Northeastern University
Boston,USA
davda.g@husky.neu.edu

1. Abstract- Twitter is one of the most popular and widely used social media platforms. It is used by people to express their opinions on eclectic subjects like politics, sports, Celebrities, etc. This research paper aims at analyzing and classifying these opinions into either positive, negative or neutral. This classification of tweets is done by using Sentiment Analysis. The Sentiment Analysis is done by extracting news from the internet based on a particular keyword. Then Tweets related to this news are analyzed. These results are then interpreted to give information regarding how people are reacting to that news in different geographical regions.

Keywords:-LogisticRegression, TfidfVectorizer, VADER Sentiment Analysis, geomap, Tweepy, AYLIEN

2. Introduction- There are various news articles present today, and every moment new ones are coming in. To see the impact of these news on people's emotion in various geographical location is of mass interest. For eg. if India wins a match against Pakistan, how are people reacting to its win, and in which geographical location this news is popular. This information is critical for companies to plan their marketing campaigns and also strategize on how to make their products and services better. Firstly, using the NewsAPI we extract the news related to the keyword we provide. For example, if we provide the keyword 'Trump' the API will fetch recent news headlines related to Trump. We then use another API called Tweepy to use these headlines as reference and then it fetches the tweets from Twitter relevant to the news articles.

These tweets will then be analyzed and classified into positive, negative or neutral. We will then visualize the analysis by displaying the different sentiments in the geo map.

3. Methods Used-

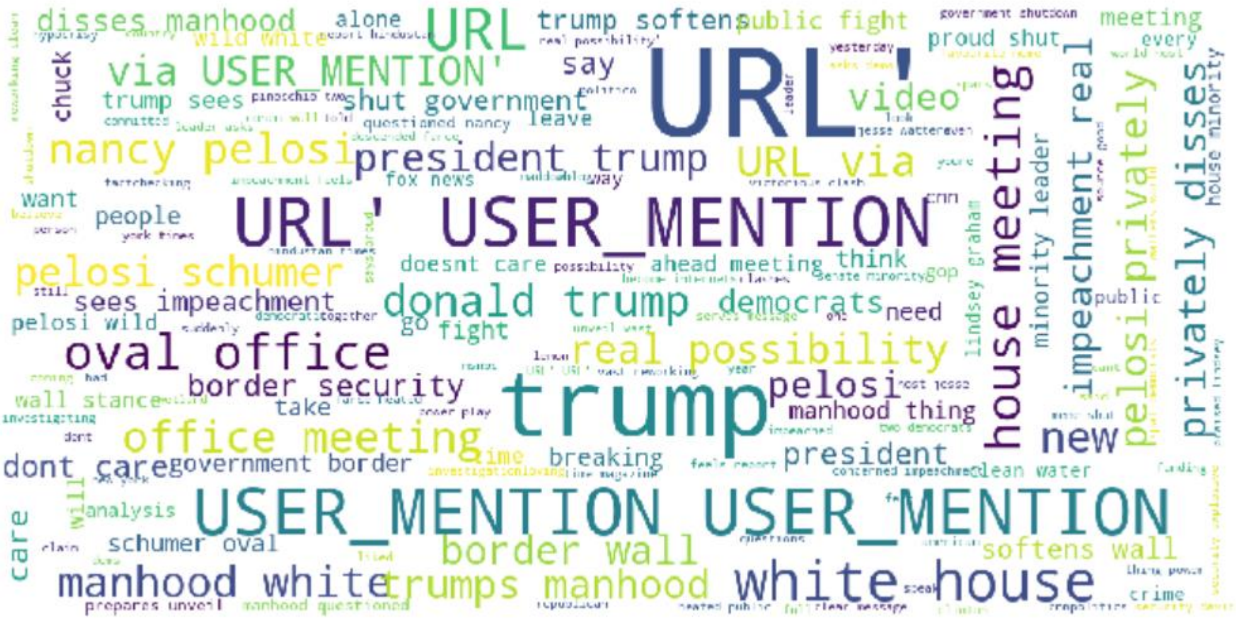
1)Training the model

Dataset which is used for training is “Sentiment140 dataset with 1.6 million tweets” and is downloaded from Kaggle. It contains 1,600,000 tweets extracted using the twitter api .The tweets have been annotated (0 = negative, 2 = neutral, 4 = positive) and they can be used to detect sentiment.

- The tweets column in the dataset is pre-processed. pre-processing consist of cleaning data such as removing multiple spaces, replacing emojis with either EMO_POS or EMO_NEG,replacing URLs with the word URL
- Below is the Word Cloud for the before Pre-processing

Vast Reworking
clear message
Reworking Clean
shut government
privately disses
Chuck Nancy
President Trump
Pelosi Schumer
Trump
message Donald
security explosive
Trump clashes
realDonaldTrump
disses Trump
House meeting
real possibility
Trump Person
Unveil Vast
leave Trump
President
Sees Impeachment
Clean Water
White House
Trump Prepares
explosive Oval
Prepares Unveil
magazine sent
Time magazine
Trump Softens
Investigation loving
Trump concerned
VIDEO
Oval Office
Office meeting
care
border wall
Nancy Pelosi
Impeachment Real
meeting
https
https co
Trump manhood
Pelosi privately
Ahead Chuck
sent clear
Donald Trump
border security
Schumer border
Person Year
Softens Wall
Water Protections
clashes Pelosi
manhood White
Trump alone
loving House
public fight
Nancy Meeting

Word Cloud after Pre-processing



- Target column in the dataset contains two values, 0 and 1. When the tweet is positive, the value in the Target is 1 and for negative tweet, the value is zero.

- The Dataset is trained using the **Logistic Regression Algorithm** which is a classification algorithm. Data is split into training data(80%) and testing data(20%).

- The training data is given to **TfidfVectorizer.**

- **TF-IDF** stands for "Term Frequency—Inverse Data Frequency".
- **Term Frequency (tf)**: gives the frequency of the word in each document in the corpus. It is the ratio of number of times the word appears in a document compared to the total number of words in that document. It increases as the number of occurrences of that word within the document increases. Each document has its own tf.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}}$$

- **Inverse Data Frequency (idf):** used to calculate the weight of rare words across all documents in the corpus. The words that occur rarely in the corpus have a high IDF score. It is given by the equation below.

$$idf(w) = log(\frac{N}{df_t})$$

Combining these two we come up with the TF-IDF score (w) for a word in a document in the corpus. It is the product of tf and idf:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

- After training the data, the model achieves an accuracy score of 78%. Using this model, predictions are made on the Live tweets later.

2) Fetching Recent News Title

After registering on the AYLIEN News API then using the app id and app key , fetch recent news headlines related to the keyword that is provided as a parameter.

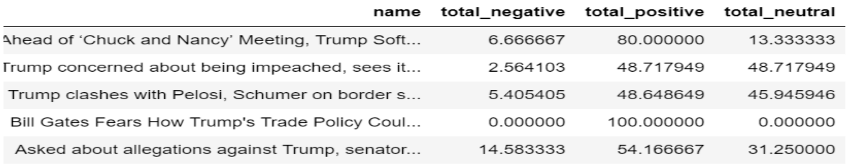
3) Collecting Tweets relevant to the News Article

The news headline is passed as a parameter to a function which returns the tweets relevant to the news headline. Tweets from Twitter are collected by using Tweepy API. The Location of the used is saved.

4)Sentiment Analysis of Tweets

Sentiment Analysis is performed using two techniques

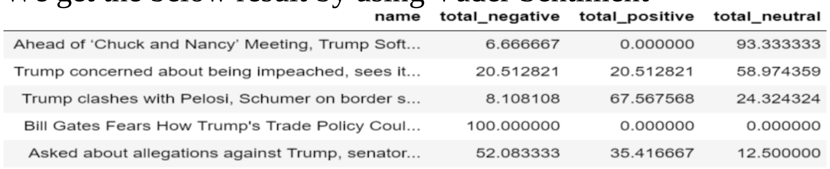
1. Logistic Regression – Predictions are made on the tweets using the Logistic Regression model which we trained earlier using the Kaggle Dataset.



name	total_negative	total_positive	total_neutral
Ahead of ‘Chuck and Nancy’ Meeting, Trump Soft...	6.666667	80.000000	13.333333
Trump concerned about being impeached, sees it...	2.564103	48.717949	48.717949
Trump clashes with Pelosi, Schumer on border s...	5.405405	48.648649	45.945946
Bill Gates Fears How Trump's Trade Policy Coul...	0.000000	100.000000	0.000000
Asked about allegations against Trump, senator...	14.583333	54.166667	31.250000

2. Vader Sentiment - VADER Sentiment Analysis. VADER (Valence Aware Dictionary and Sentiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media.

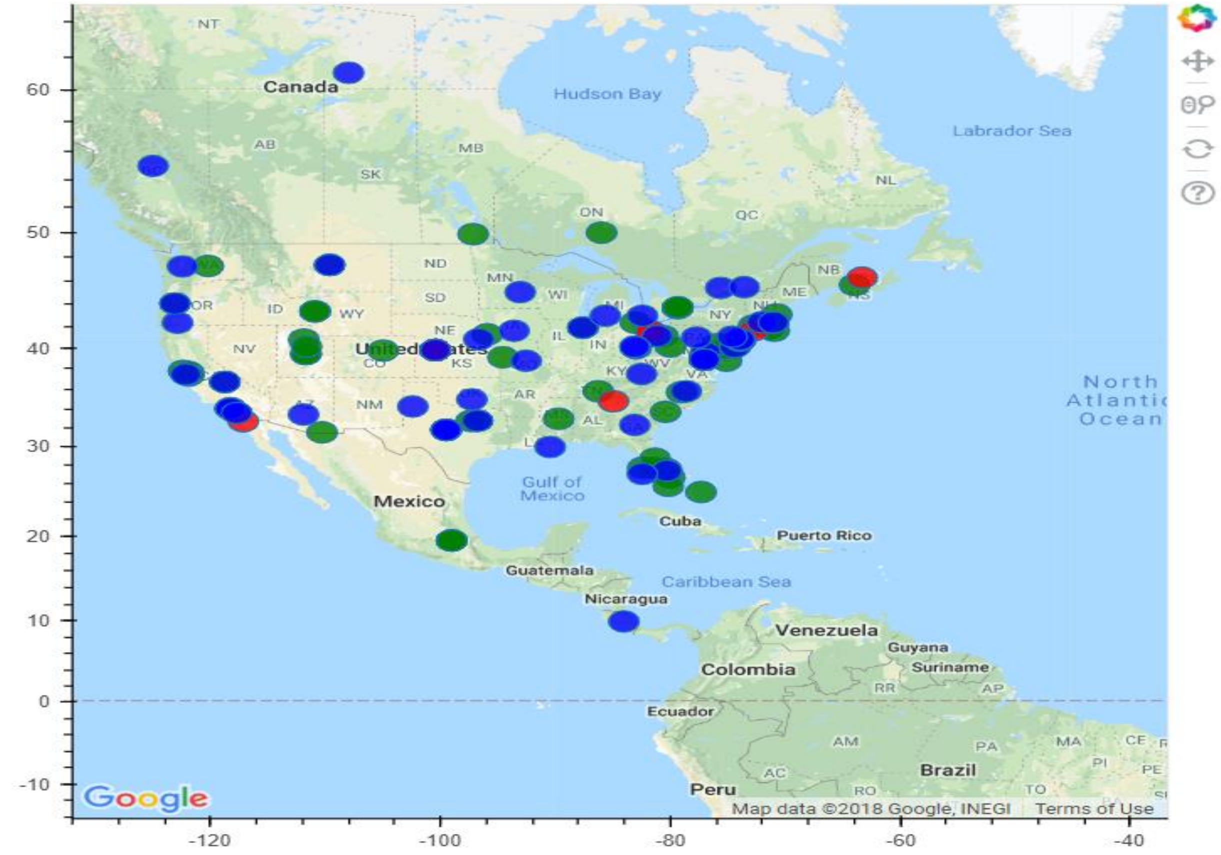
We get the below result by using Vader Sentiment



name	total_negative	total_positive	total_neutral
Ahead of 'Chuck and Nancy' Meeting, Trump Soft...	6.666667	0.000000	93.333333
Trump concerned about being impeached, sees it...	20.512821	20.512821	58.974359
Trump clashes with Pelosi, Schumer on border s...	8.108108	67.567568	24.324324
Bill Gates Fears How Trump's Trade Policy Coul...	100.000000	0.000000	0.000000
Asked about allegations against Trump, senator...	52.083333	35.416667	12.500000

4.Results-

Result of Logistic Regression model and the result of Vader Sentiment Analysis is compared. By using the Logistic Regression, accuracy of 78% is achieved which is good.
Using Geopi, the latitude and longitude of the users location are retrieved .
The Sentiment of tweets is Plotted in World Map by using the latitude and longitude of users location.
To distinguish between the sentiments, the output are divided into different colors
Green for Positive Sentiment, Red for Negative Sentiment and Blue for Neutral Sentiment.
Below is the Bokeh Plot for the recent News Article Title relating to the Keyword Trump.
Tweets about the News Title containing "Trump" are fetched and their Sentiment Analysis using the Logistic Regression model is evaluated.



5.Discussion-

Dataset used from the Kaggle contained 1.6 million tweets. The model was trained using this dataset by Logistic Regression. From the AYLIEN News API, recent news headlines related to the keyword is fetched. The tweets relevant to the news headline are Collected. Tweets from Twitter are collected by using Tweepy API. The Location of the user is saved. The sentiment of users is plotted in Bokeh plot. By going through the Bokeh plot, one can evaluate the reaction(Positive, Negative and Neutral) of users in different geographical regions and this can be used in Marketing.

6.References

http://docs.tweepy.org/en/v3.5.0/getting_started.html

https://newsapi.org/

https://developers.google.com/maps/documentation/javascript/get-api-key

https://bokeh.pydata.org/en/latest/docs/user_guide.html#userguide

https://www.nltk.org/api/nltk.sentiment.html

https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

https://www.kaggle.com/nikhilsable/sentiment-using-airline-tweets-using-vader

https://www.journaldev.com/19527/bokeh-python-data-visualization

http://vprusso.github.io/blog/2018/natural-language-processing-python-3/

https://www.kaggle.com/kazanova/sentiment140/home

https://medium.freecodecamp.org/how-to-process-textual-data-using-tf-idf-in-python-cd2bbc0a94a3

https://github.com/cjhutto/vaderSentiment