TWITTER SENTIMENTAL ANALYSIS AND VISUALISATION

By: Srijan Malhotra (SM9439/N18390405)

ABSTRACT:

Twitter is one of the leading social networking platforms that has been attracting a huge amount of attention not only from general users but also from researchers all around the world. A large amount of data is being generated almost every day which can be used to analyze and visualize the usage of social media data which not only helps understand how it is being used but also allows helps understand the general trend for that specific period. [1]

In this project, I have made use of NLP techniques that help me in extracting the emotions which in turn helps in analyzing, visualizing, and providing a sentimental analysis for the Twitter data set and then providing a conclusion that would allow the reader to understand the trend of the entire data set.

1. INTRODUCTION:

Today social media not only acts as a medium to communicate but is also the leading method of data consumption.[2] Twitter being a leader in its field allows people to communicate their thoughts without any barriers. It not only helps connect friends and family but has been involved very actively in various events as sponsors etc. It has been used at several events and the tweets indicate what the people think about the event and if they support it or not.

This being said many patterns emerge from every single day's data and to understand that people need to go through the data which is very tedious and taxing. To avoid that I have focused on analyzing and providing a few visualizations that best help understand the data set and help someone save time instead of going through a plethora of data.

Most of the research until now has focused on providing basic visualizations but my goal has been to provide the users with a much more enhanced and detailed description like the location and frequency of tweets, finding the polarity and use that to not only make a graph but also make a word cloud and then further segregate that into 2 different word clouds and finally also analyze the retweets along with the date and time all of this, in the end, would also lead towards a sentimental analysis of the Twitter data set.

2. PROBLEM STATEMENT:

Analyze Twitter Data Set to provide visualizations and a word cloud that depicts the sentimental analysis of the data.

3. METHODOLOGY AND VISUALISATIONS:

To achieve the ultimate goal of helping the user understand and fully absorb the data I had to follow some steps. These steps not only allowed me to fetch the data but also allowed me to clean the data, remove unnecessary and unwanted parts of the data. Furthermore, that entitled me to use NLP techniques on the dataset, then plot individual visualizations and the word cloud for sentimental analysis.
Following are the detailed description of each step followed by me in this project:

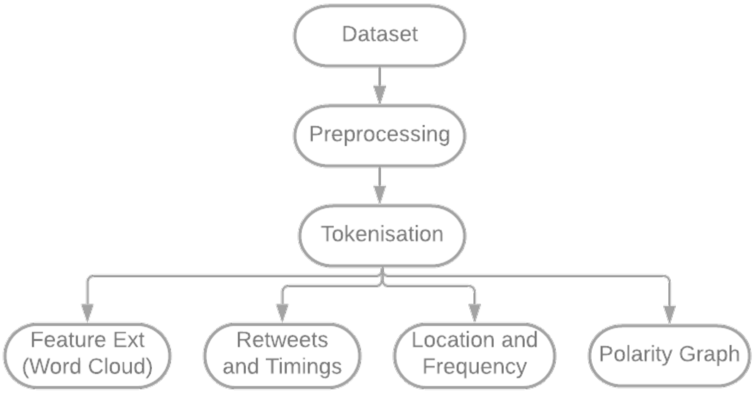

Fig 1. Step by Step Description

3.1. Preprocessing:

I have made use of Tweepy which is an open-sourced python library to access Twitter API. Data from Twitter API is used to perform sentiment analysis. A series of preprocessing steps need to be applied to this data before we can start analyzing and visualizing it. A query is passed which helps in gathering the data and then it is collected using the Streaming API. The data received will be converted according to the needs (for instance, pandas data frame is used here).

This includes:

- Converting upper case to lower case
- Removal of URL
- Removal of Stopwords
- Removal of Handles, Emoticons
- Removing repeated characters
- Stemming: - It is used to replace words with their root words to reduce different words to the same type.
- Parts of speech: - Assigning a tag to each word in the text and classifying it as a verb, adjective, etc.
- Tokenization: - Converts meaningful pieces of data into a random string of characters. (POS Tagging)

3.2. Feature Extraction: -

To achieve feature extraction using word cloud we make use of the polarity score that has been calculated and appended in the data set. Based on this score we can classify tweets as positive, negative, or neutral. Through this classification, I have come up with three word clouds wherein the first one is the word cloud for the entire dataset, the second and the third respectively act as word clouds for positive and negative tweets. This helps in understanding the general trend and how people react to it.

Ways to classify: -

- **Bag of Words: -**
  All the tokens in the data set are used a vocabulary. The frequency with which the token appears in the vocabulary of every document consists of the count of every term, the count of documents with tokens which in turn would determine the reverse frequency.

- **Word doc2vec: -**
  Doc2vec uses a three-layer shallow deep neural network to measure the context of a document and connect similar context phrases. This is not a monolithic algorithm. There are two different variations, SKIP GRAM and CBOW. This is being used to train a large data set like Twitter so that it produces meaningful embeddings which lead to amazing feature extractions.

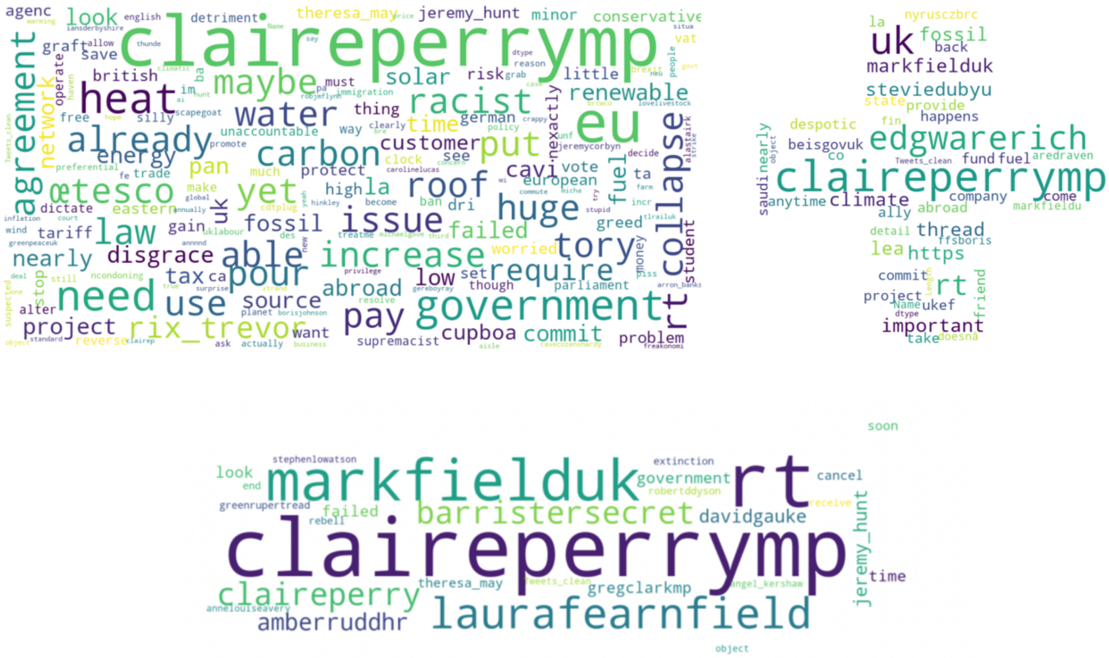

Fig 2. Feature and Sentiment Analysis using Word Cloud

3.3. <u>Maximum Retweets and Timings: -</u>

This makes use of a one-dimensional labeled array that is capable of holding any type of data. In this case, we use pd.Series which sorts data in a 1D Array and then plots it according to the requirement. That being retweets and the date and time the tweet was retweeted the most.

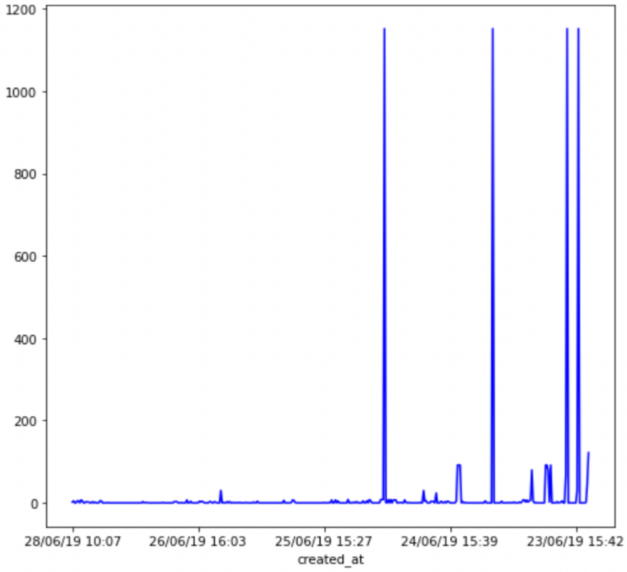

Fig 3. Max Retweets along with Date and Time

3.4. <u>Location and Frequency of Tweets: -</u>

This is used to fetch the number of tweets and the location that they were made from. This graph allows the user to understand the demography from where people are most active and the number of times they tweet about a given topic.

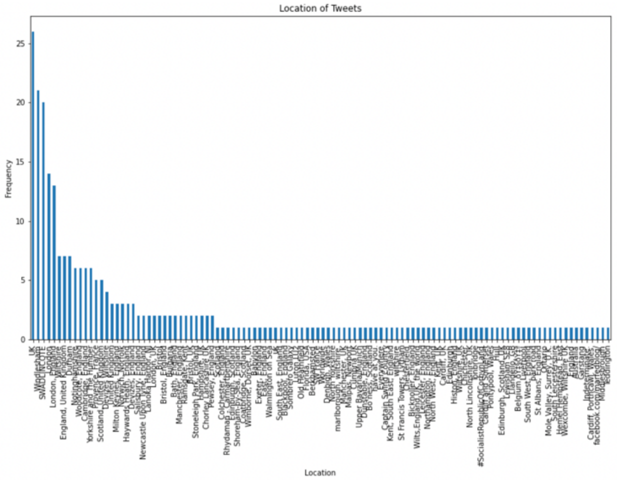

Fig 4. Max Freq of Tweets from different Locations

3.5. Polarity Comparison graph for Positive and Negative Tweets: -

This enables the user to understand the impact of the tweet by differentiating it between positive or negative replies. This would not only help a user understand the different types of tweets but would also help him get the current trend.

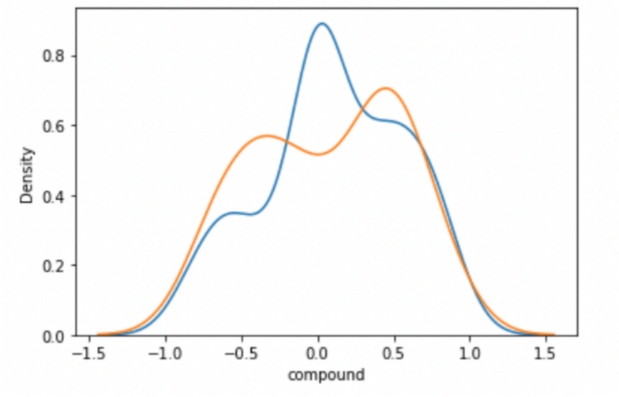

Fig 5. Polarity Comparison

4. CONCLUSION: -

In conclusion, analysis of data can be highly simplified by visualizing it as that not only keeps the user engaged but also allows him to absorb the data. In this project, I have made use of Python, NLP techniques, and Libraries to simplify the Twitter data of a single user for the people to grasp. Furthermore, this shows how the data can be used to understand current trend or people's behavior, interests in general, or about something happening on Twitter and this could help someone make their analysis.[1]

I had kept the first paper in the references section as my reference and I have managed to complete most of it except the last part (Streamgraph) since the data I am using is categorical and the data expected for that is numerical.

Finally, I have achieved sentimental analysis through the polarity for positive, negative, and neutral tweets and have also compared maximum retweets along with the locations that give out the maximum tweets. This would help the user get an understanding about which location has a better standing and what is the level of sentiments for the topic in that location.

5. REFERENCES: -

1. Stojanovski, D., Dimitrovski, I., & Madjarov, G. (2014). Tweetviz: Twitter data visualization. *Proceedings of the data mining and data warehouses*, *1*(2).

2. Sechelea, A., Do Huu, T., Zimos, E., & Deligiannis, N. (2016, May). Twitter data clustering and visualization. In *2016 23rd international conference on telecommunications (ICT)* (pp. 1-5). IEEE.

3. Bhulai, S., Kampstra, P., Kooiman, L., Koole, G., Deurloo, M., & Kok, B. (2012). Trend visualization on Twitter: what’s hot and what’s not?. *Data analytics*, 43-48.

4. https://analyticsindiamag.com/complete-tutorial-on-text-preprocessing-in-nlp/

5. https://amueller.github.io/word_cloud/auto_examples/masked.html