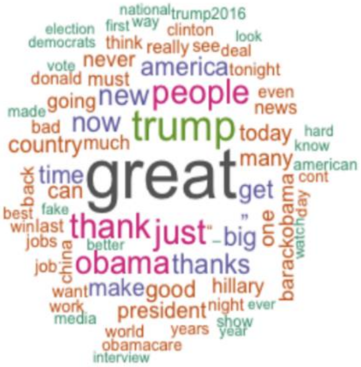


SENTIMENT ANALYSIS: TRUMP TWEETS POSITIVE OR NEGATIVE?

IDS 561 – BIG DATA ANALYTICS – FINAL PROJECT

Group 15:
Aditya Gajula – 660252623
Ritu Gangwal – 670646774
Viharika Bharti - 655974244

SENTIMENT ANALYSIS: TRUMP TWEETS POSITIVE OR NEGATIVE?

Problem Setting:

Sentiment analysis is the process of ‘computationally’ determining whether a piece of writing is positive, negative, or neutral. It is also known as opinion mining, deriving the opinion or attitude of a speaker. Sentiment analysis is used by several data analytics companies on a variety of subjects. The US Presidential Elections are held on the 3rd of November 2020 and with the results almost out, this is a good time to do an overall study and understand how the public mood and opinion mattered in different US states towards the candidate — Republican current president Donald Trump.

Twitter boasts 330 million monthly active users, which allows businesses to reach a broad audience and connect with customers without intermediaries. On the downside, there’s so much information that it is hard for companies to quickly detect negative social mentions that could impact any social event. Therefore social listening, which involves monitoring tweets on Twitter platforms, has become a key strategy in sentiment analysis.

The overall benefits of Twitter sentiment analysis include:

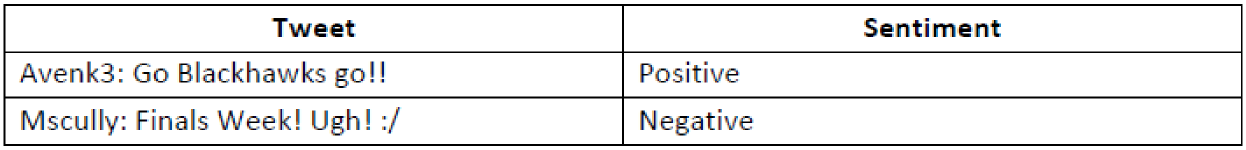
- Scalability: Analyze hundreds or thousands of tweets mentioning your brand and automate manual tasks. Easily scale sentiment analysis tools as your data grows and gain valuable insights on the go.
- Real-Time Analysis: Twitter sentiment analysis is essential for monitoring sudden shifts in customer moods, detecting if complaints are on the rise, and for taking action before problems escalate. With sentiment analysis, monitor brand mentions on Twitter in real-time and gain actionable insights.
- Consistent Criteria: Avoid inconsistencies that stem from human error. Customer reps won’t always agree on which tag to use for each piece of data, so you may end up with inaccurate results. Instead, machine learning models perform sentiment analysis using one set of rules, so you can ensure all your Twitter data is tagged consistently.

This project aims to do sentiment analysis of political views of Twitter users during the 2020 American Presidential Election campaign. This is an important study as public opinion for a particular candidate will impact the potential leader of the country. We are relying on Twitter as it acquires a large diverse data set representing the current public opinion about President Trump.

The business objective of this project is to analyze the following questions:

- What is the public sentiment distribution across the various locations/ states in the U.S.?
- Which states require more strategic campaigning/ awareness by the Trump Government?
- Does the positivity of a tweet influence the quantity of retweets/favorites? We performed the analysis to study the current state of public opinions.

To this project, we identified sentiment conveyed in a given tweet as “positive” or “negative”. Shown below are a couple of examples:



Tweet	Sentiment
Avenk3: Go Blackhawks go!!	Positive
Mscully: Finals Week! Ugh! :/	Negative

<u>Challenges:</u>

Tweets are limited to only 140 characters per tweet and hence are very short. It can be something as simple as an acronym or as a headline. The language used can be very casual or with improper grammar, repeated words. To mine through these categories of text and get an understanding of the underlying sentiment was a challenging task.

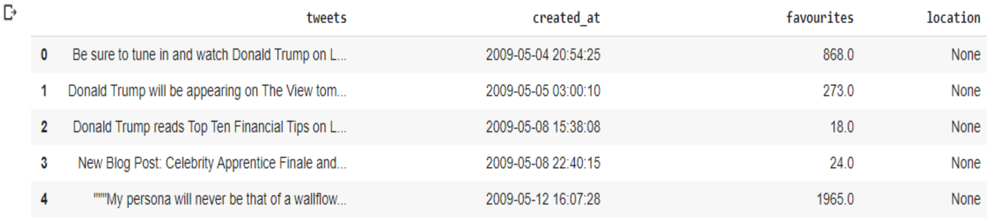
<u>Data Description:</u>

For this analysis, tweets were collected from multiple sources and the time considered was one year. **931,631** tweets were collected using Kaggle (historical datasets), Twitter API (recent tweets) and Twint API. Since the data is collected from multiple sources, the preliminary data exploration included getting the uniform data from all sources with the same number of variables to be considered for analysis. The variables considered are:

- Tweets - the tweet content posted by a user
- Tweet created date - date when the tweet is created
- favourites/tweets liked - number of likes the tweet received
- Location - user's given location in the profile
- User screen name - the user's screen name on the twitter platform.

Each variable considered for the analysis has a specific objective. "Tweets" is the main variable we are dealing with for our sentiment analysis, where we get the actual text/words of the tweet content. The other variables are considered as additional metrics needed for our project outcomes.

Initial Dataframe:



	tweets	created_at	favourites	location
0	Be sure to tune in and watch Donald Trump on L...	2009-05-04 20:54:25	868.0	None
1	Donald Trump will be appearing on The View tom...	2009-05-05 03:00:10	273.0	None
2	Donald Trump reads Top Ten Financial Tips on L...	2009-05-08 15:38:08	18.0	None
3	New Blog Post: Celebrity Apprentice Finale and...	2009-05-08 22:40:15	24.0	None
4	"""My persona will never be that of a wallflow...	2009-05-12 16:07:28	1965.0	None

Figure 1. Initial Dataframe

<u>Basic project pipeline:</u>

The below figure explains the basic pipeline of our project which shows how the data flow from original source to its output for visualizations.

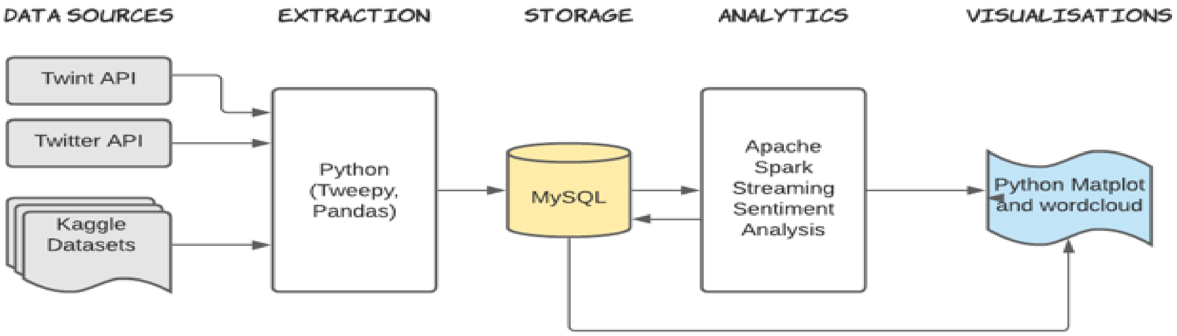


Figure 2. Project pipeline

Techniques:

As the data was collected from multiple sources, the need for a database was crucial and the project uses MySQL database to store and access the data due to its structured property. **MySQL** is a relational database management system based on SQL – Structured Query Language. The application is used for a wide range of purposes, including data warehousing, e-commerce, and logging applications. We have used python and twitter API to directly store the tweet data into MySQL.

We created a database called 'tweetsschema' using MySQL workbench with a table called 'trumptweets'. Columns used for extracting tweets are shown below:

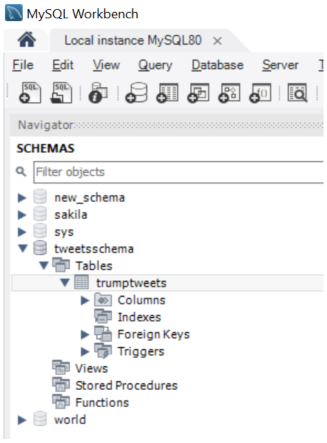


Figure 3. Database schema

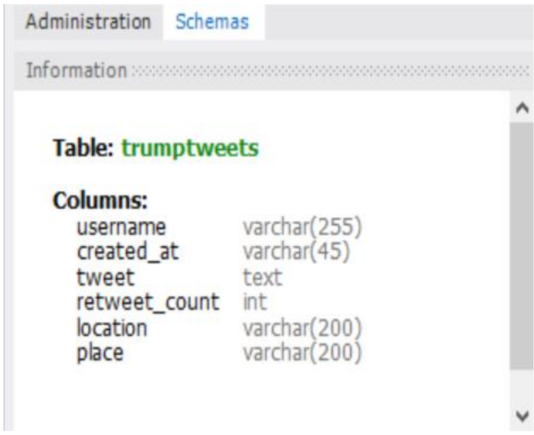


Table: trumptweets

Columns:

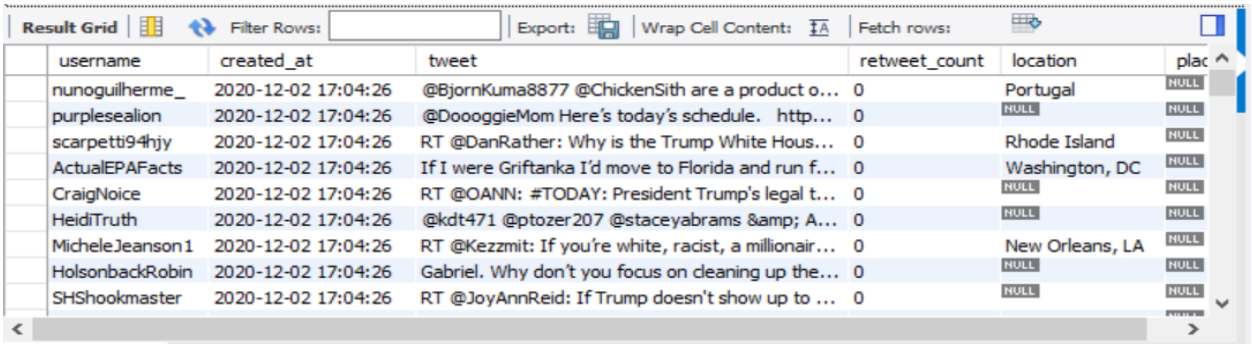
username	varchar(255)
created_at	varchar(45)
tweet	text
retweet_count	int
location	varchar(200)
place	varchar(200)

Figure 4. Table schema

After setting up the database, we followed the below steps to establish connection to Twitter API:

- We used Tweepy library which made it very easy for us to connect to the API and start streaming the data.
- We have set up our tokens and secret keys and our password for the database.
- Created a class inheriting from StreamListener.
- Instantiate an object from this class.
- Use this object to connect to the API.

After connection is established, data is directly collected in the database table.



username	created_at	tweet	retweet_count	location	plac
nunoguilherme_	2020-12-02 17:04:26	@BjornKuma8877 @ChickenSith are a product o...	0	Portugal	NULL
purplesealion	2020-12-02 17:04:26	@DoooggieMom Here's today's schedule. http...	0	NULL	NULL
scarpetti94hjy	2020-12-02 17:04:26	RT @DanRather: Why is the Trump White Hous...	0	Rhode Island	NULL
ActualEPAFacts	2020-12-02 17:04:26	If I were Griftanka I'd move to Florida and run f...	0	Washington, DC	NULL
CraigNoice	2020-12-02 17:04:26	RT @OANN: #TODAY: President Trump's legal t...	0	NULL	NULL
HeidiTruth	2020-12-02 17:04:26	@kdt471 @ptozer207 @staceyabrams &amp; A...	0	NULL	NULL
MicheleJeanson1	2020-12-02 17:04:26	RT @Kezzmit: If you're white, racist, a millionair...	0	New Orleans, LA	NULL
HolsonbackRobin	2020-12-02 17:04:26	Gabriel. Why don't you focus on cleaning up the...	0	NULL	NULL
SHShookmaster	2020-12-02 17:04:26	RT @JoyAnnReid: If Trump doesn't show up to ...	0	NULL	NULL

Figure 5. Extracted tweets to MySQL

Twitter API: The Twitter API enables programmatic access to Twitter in unique and advanced ways. We used it to analyze, learn from, and interact with Tweets, users, and other key Twitter resources. In order to use the Twitter API, we need to have a Twitter Developer Account which gives us access to an access token, an access token secret, an api key and an api secret key. We used python's configparser to read those data in order to use the twitter API.

Twitter uses pagination a lot in their API development. In order to perform pagination, Twitter supplies a page/cursor parameter with each of their requests. As a result, this requires a lot of boilerplate code just to manage the pagination loop. To help make pagination easier and require less code the Twitter API or Tweepy has the Cursor object.

Now, since the Cursor is passed callable, we can not pass the parameters directly into the method. Instead the parameters are passed into the Cursor constructor method. For our api to search tweets, we pass the following parameters:

1. **q:** This is the keyword to be searched in the tweet. For our project, we pass the name of the candidate. (#trump)
2. **lang:** This is the language of the tweets we want to retrieve from the API. Since the USA is largely an English-speaking country with English also being the official language, we retrieve tweets made in English.
3. **Since:** This is the date from which we want to retrieve tweets. For our purposes, we are interested in the current political discourse to understand the current mood ahead of the elections.

Another factor which needs to be kept in mind is the tweepy search api retrieves a maximum of 1500 tweets at a time followed by a cooldown period of 15 minutes.

Twint API: Twint is an advanced Twitter scraping tool written in Python that allows for scraping Tweets from Twitter profiles without using Twitter's API. Twint utilizes Twitter's search operators to let you scrape Tweets from specific users, scrape Tweets relating to certain topics, hashtags & trends, or sort out sensitive information from Tweets like email and phone numbers. I find this very useful, and you can get really creative with it too. Twint also makes special queries to Twitter allowing you to also scrape a Twitter user's followers, Tweets a user has liked, and who they follow without any authentication, API, Selenium, or browser emulation.

Data Pre-processing:

Before analysing the tweets, some basic data cleaning/pre-processing steps followed are:

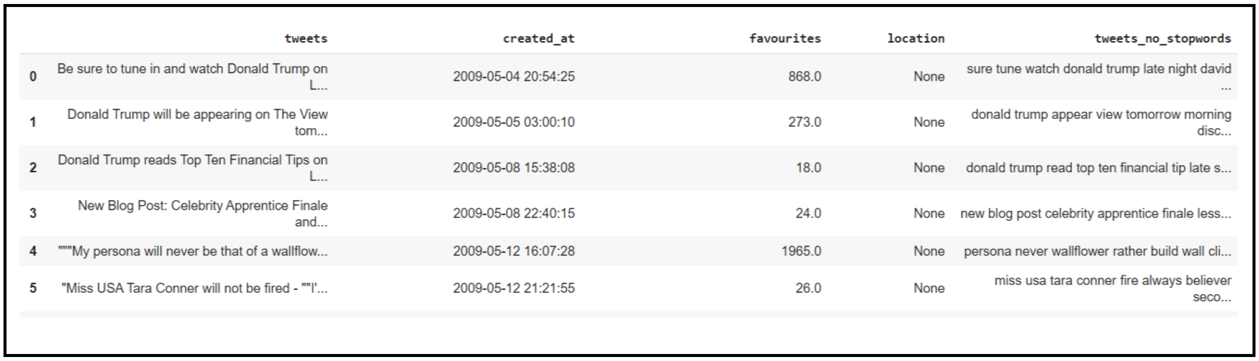
- HTML decoding of the data is performed using Beautiful Soup.
- The tokenizer correctly handled URLs, common emoticons, phone numbers, twitter mentions and hashtags, repetition of symbols.
- Remove the stop words, perform Parts of speech (POS) tagging, lemmatizing.
- Abbreviations were replaced by their full form and emoticons such as ":)" were replaced by named tags, e.g. <happy>.

For replacing the acronyms, we have formulated an acronym dictionary from all the slangs listed in www.noslang.com/dictionary/. This dictionary has translations for all the 5300 acronyms.

For handling the presence of emoticons, we have compiled a list of 160 emoticons listed in Wikipedia (http://en.wikipedia.org/wiki/List_of_emoticons) and have assigned them polarity according to their usage. For ex: “:)” this smile symbol has been assigned a positive polarity as it is mostly used for expressing happiness.

We have replaced negation words with 'NOT' and for choosing negations we have referred to the negative words that are provided by the Grammarly handbook Our algorithm converts can't to can NOT (Since, n't is the negation here) and cannot to NOT. (www.grammarly.com/handbook/sentences/negatives/).

After performing all data pre-processing steps, our data-frame output is shown below:



	tweets	created_at	favourites	location	tweets_no_stopwords
0	Be sure to tune in and watch Donald Trump on L...	2009-05-04 20:54:25	868.0	None	sure tune watch donald trump late night david ...
1	Donald Trump will be appearing on The View tom...	2009-05-05 03:00:10	273.0	None	donald trump appear view tomorrow morning disc...
2	Donald Trump reads Top Ten Financial Tips on L...	2009-05-08 15:38:08	18.0	None	donald trump read top ten financial tip late s...
3	New Blog Post: Celebrity Apprentice Finale and...	2009-05-08 22:40:15	24.0	None	new blog post celebrity apprentice finale less...
4	"""My persona will never be that of a wallflow...	2009-05-12 16:07:28	1965.0	None	persona never wallflower rather build wall cli...
5	"Miss USA Tara Conner will not be fired - ""I'...	2009-05-12 21:21:55	26.0	None	miss usa tara conner fire always believer seco...

Figure 6. Dataframe after data pre-processing

Sentiment Analysis:

Sentiment analysis is done using three different methods and then we have compared the results by putting all the results into the same dataframe.

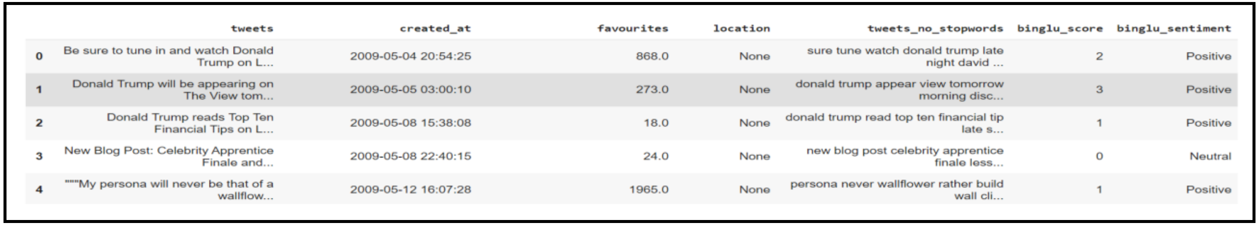
1. Sentiment analysis using bing lu systems
2. Sentiment analysis through textblob
3. Sentiment analysis through Vader

1. Sentiment score using Bing Lu Dictionary:

One of the methods used to calculate sentiment score is comparing the tweet content with positive and negative words from Bing Lu dictionary. Detailed steps are explained below:

- We have taken positive and negative words from the dictionary.
- Assigned value -1 to all negative words and +1 to positive words.
- Words that are not matching with dictionary words are not taken into account.
- Lastly, we compared the tweet words with the dictionary's positive and negative words and then calculated sentiment score by adding and subtracting word value.

Output:



	tweets	created_at	favourites	location	tweets_no_stopwords	binglu_score	binglu_sentiment
0	Be sure to tune in and watch Donald Trump on L...	2009-05-04 20:54:25	868.0	None	sure tune watch donald trump late night david ...	2	Positive
1	Donald Trump will be appearing on The View tom...	2009-05-05 03:00:10	273.0	None	donald trump appear view tomorrow morning disc...	3	Positive
2	Donald Trump reads Top Ten Financial Tips on L...	2009-05-08 15:38:08	18.0	None	donald trump read top ten financial tip late s...	1	Positive
3	New Blog Post: Celebrity Apprentice Finale and...	2009-05-08 22:40:15	24.0	None	new blog post celebrity apprentice finale less...	0	Neutral
4	"""My persona will never be that of a wallflow...	2009-05-12 16:07:28	1965.0	None	persona never wallflower rather build wall cli...	1	Positive

Figure 7. Output from Bing Lu sentiment analysis

2. Sentiment analysis through textblob:

The textblob package approach required pre-defines set of categorized words which can be downloaded from the NLTK database. This is a Python (2 or 3) library which is used to process text data. It contains simple APIs to perform NLP tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification etc. We will be using this library to perform sentiment analysis. This contains two major metrics –

i) Polarity – The polarity value varies from -1 and 1, where -1 denotes negative sentiment and 1 shows positive sentiment.

- if polarity score < 0 -> 'Negative'
- else if polarity score == 0 -> 'Neutral'
- else -> 'Positive'

ii) Subjectivity – This score tells us if the text is more of an opinion or a fact. The value for subjectivity score varies from 0 to 1 where 0 means objective (fact) and 1 means subjective (opinions).

Cleaned data is fed into a python model with a textBlob library which will give us the output in the form of polarity and sensitivity.

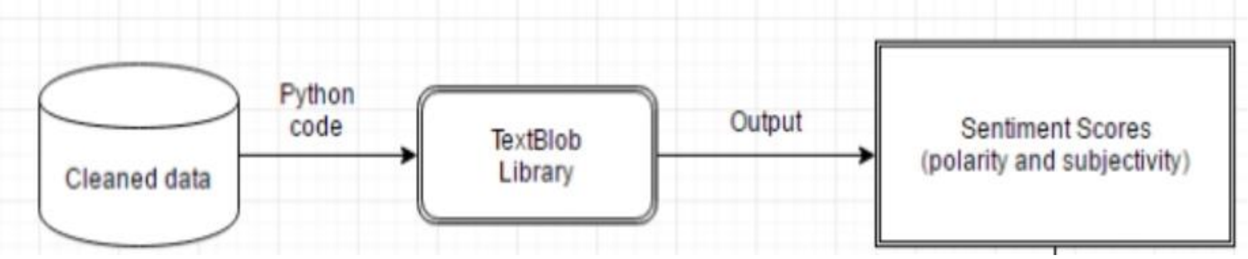
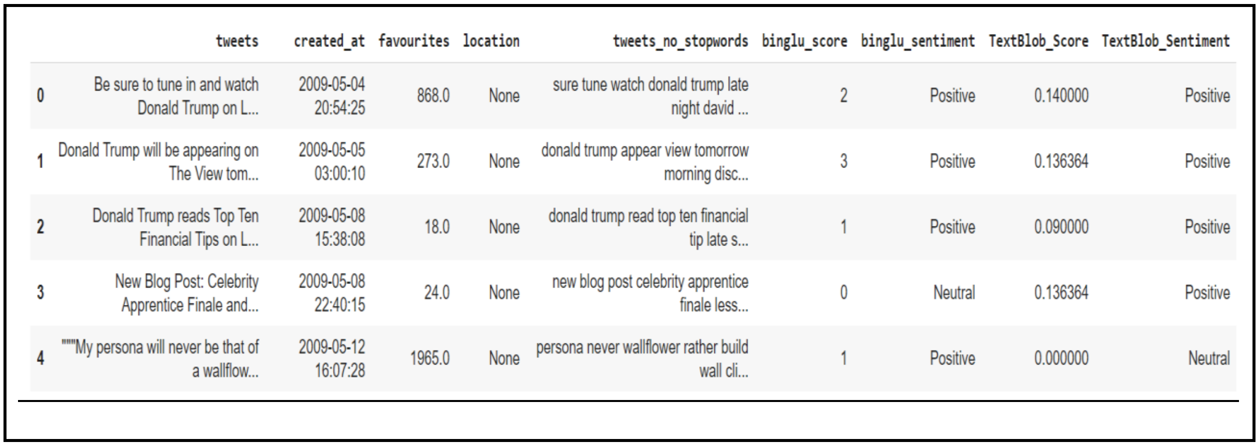


Figure 8. TextBlob sentiment analysis

TextBlob's output for a polarity task ranges from [-1,1] where -1 is a negative polarity and +1 is a positive polarity. Score can also be 0 which stands for neutral polarity.

<u>Output:</u>



	tweets	created_at	favourites	location	tweets_no_stopwords	binglu_score	binglu_sentiment	TextBlob_Score	TextBlob_Sentiment
0	Be sure to tune in and watch Donald Trump on L...	2009-05-04 20:54:25	868.0	None	sure tune watch donald trump late night david ...	2	Positive	0.140000	Positive
1	Donald Trump will be appearing on The View tom...	2009-05-05 03:00:10	273.0	None	donald trump appear view tomorrow morning disc...	3	Positive	0.136364	Positive
2	Donald Trump reads Top Ten Financial Tips on L...	2009-05-08 15:38:08	18.0	None	donald trump read top ten financial tip late s...	1	Positive	0.090000	Positive
3	New Blog Post: Celebrity Apprentice Finale and...	2009-05-08 22:40:15	24.0	None	new blog post celebrity apprentice finale less...	0	Neutral	0.136364	Positive
4	"""My persona will never be that of a wallflow...	2009-05-12 16:07:28	1965.0	None	persona never wallflower rather build wall cli...	1	Positive	0.000000	Neutral

Figure 9. Output from TextBlob sentiment analysis

3. Sentiment analysis usingVader:

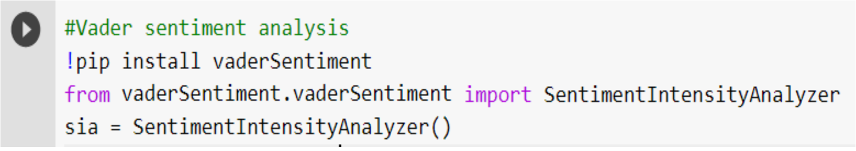
VADER (Valence Aware Dictionary and Sentiment Reasoner) is a lexicon and rule-based tool that is attuned to sentiments expressed on social media. VADER has been found to work well with social media texts, product and movie reviews etc. It tells us how positive and negative the sentiment is. It is fully open sourced under the MIT License. To get the sentiment score from VADER, a function called “polarity_scores()” is used to get different polarity scores.

- Positive
- Neutral
- Negative
- Compound

Positive, Neutral and Negative scores tell us what proportion of text falls in which category. The sum of all the scores should add up to 1.
Compound score calculates the sum of all ratings and normalized between -1 to 1. Most negative score is denoted by -1 and most positive score is denoted by 1.
The compound score is compared with the threshold value to categorize it Positive, Negative or Neutral sentiment. We used the value 0.05 and below is the algorithm –

- Compound value >= 0.05 = Positive
- Compound value <= -0.05 = Negative
- (Compound value > -0.05) and (Compound value < 0.05) = Neutral

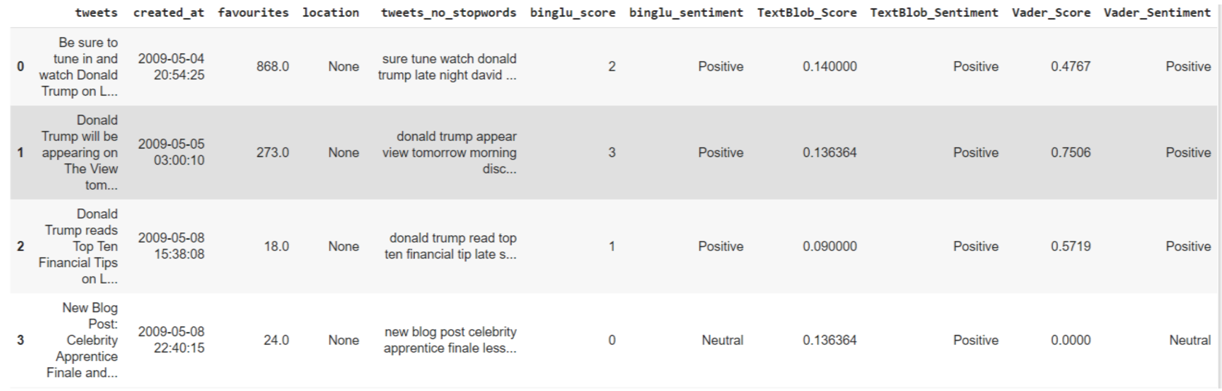
Vader sentiment can be simply installed in python using the ‘**!pip install vaderSentiment**’ statement. Once vader is installed, we can call the analyzer, **SentimentIntensityAnalyzer()** that takes in a string and returns a dictionary of scores of categories- positive, negative or neutral.



```
#Vader sentiment analysis
!pip install vaderSentiment
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
sia = SentimentIntensityAnalyzer()
```

Figure 10. Vader installation

Output:



	tweets	created_at	favourites	location	tweets_no_stopwords	binglu_score	binglu_sentiment	TextBlob_Score	TextBlob_Sentiment	Vader_Score	Vader_Sentiment
0	Be sure to tune in and watch Donald Trump on L...	2009-05-04 20:54:25	868.0	None	sure tune watch donald trump late night david ...	2	Positive	0.140000	Positive	0.4767	Positive
1	Donald Trump will be appearing on The View tom...	2009-05-05 03:00:10	273.0	None	donald trump appear view tomorrow morning disc...	3	Positive	0.136364	Positive	0.7506	Positive
2	Donald Trump reads Top Ten Financial Tips on L...	2009-05-08 15:38:08	18.0	None	donald trump read top ten financial tip late s...	1	Positive	0.090000	Positive	0.5719	Positive
3	New Blog Post: Celebrity Apprentice Finale and...	2009-05-08 22:40:15	24.0	None	new blog post celebrity apprentice finale less...	0	Neutral	0.136364	Positive	0.0000	Neutral

Figure 11. Output from Vader sentiment analysis

Data analysis on resulting data frame:

1. We have used Word Cloud for data visualization. This technique is used for representing text data in which the size of each word indicates its frequency or importance. We can see from the below world cloud that most frequent tweets contain words like realDonaldTrump, Trump, corrupt, comments etc.

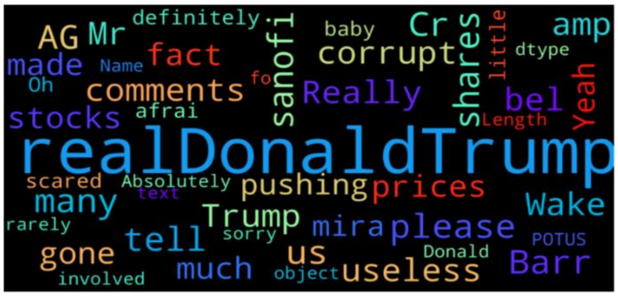


Figure 12. Word cloud

2. Now using textBlob output, we did count analysis on resulting positive, negative or neutral tweets. As we see, more than 7,00,000 tweets are categorized as neutral. Positive tweets are more in number when compared with negative tweets.

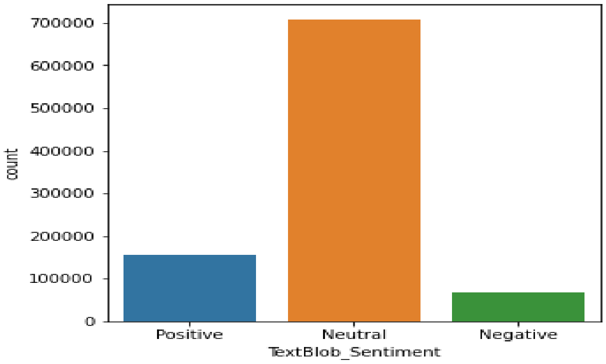


Figure 13. Tweet sentiment count

3. Our dataset contains data from 12 years starting from 2009 to 2020. Since 2020 was the year of election, and the Trump government was campaigning for re-election, we can see the number of tweets are very high compared to previous years. There could be other reasons as well for such a high number of tweets this year such as increase in usage of social media platforms by people to express their emotions and views and different social events.

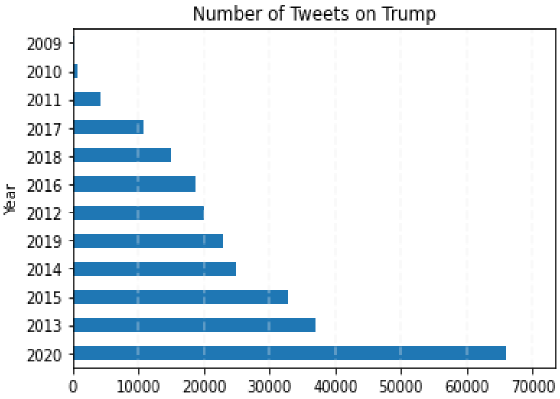


Figure 14. Number of tweets per year

4. Time series analysis for year 2020 as per text blob: From below plot, we can see how sentiment of people changed throughout the year(2020) from January to December. During COVID-19 and before the election, we see that there is a drastic decrease in positive tweets for Trump. There were many social events like COVID-19 pandemic, BlackLivesMatter movement happening during this year which could affect the tweet sentiment towards the Government. After the election, we see that the tweets show less change in trend.

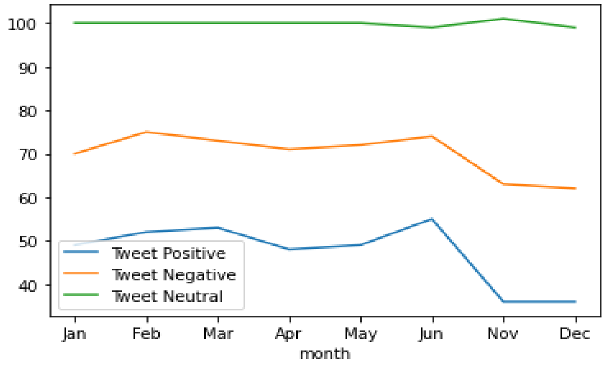


Figure 15. Time series analysis

Location based analysis:

Social media platforms such as Twitter have not only changed the way we interact with each other but also changed the way of sharing our comments and views on world events.

Hence, after collecting the data, processing it and calculating sentiment score, we decided to do location-based analysis on Trump's tweets sentiment which would help their Government to strategize their campaigning.

The side graph shows a huge shortcoming of our analysis that our dataset is not well distributed. This means we have some states like California, New York, Florida, Texas, Washington which have more data than others.
Now, to proceed with our analysis, we followed below approach:

- In our dataset, there are many rows for which, location field is empty or none. For location based analysis, all these rows are dropped.
- We have taken a textBlob sentiment score for this analysis.
- Number of positive, negative and neutral tweets are calculated for each state in the USA which gives us the conclusion if the state has more supporters or haters.

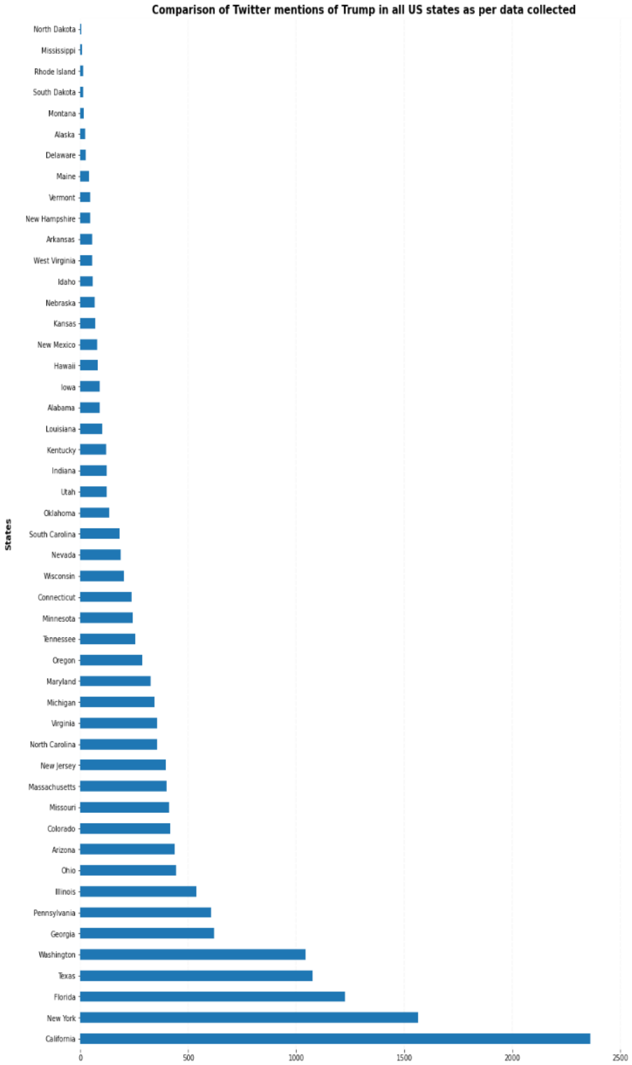


Figure 16. State wise data distribution

Below plot shows which state had more positive, negative or neutral sentiment towards Trump.

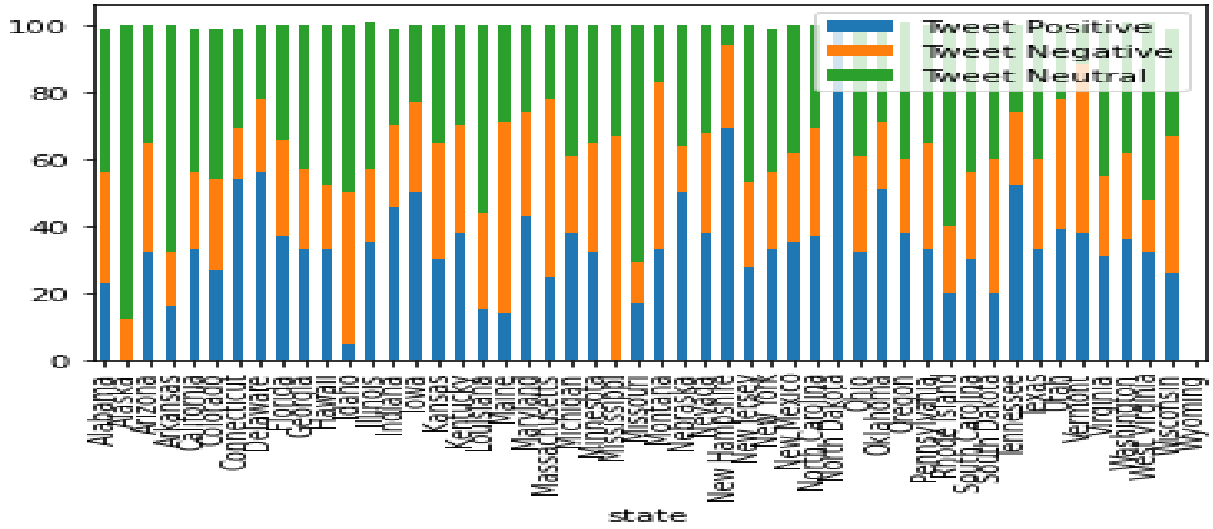
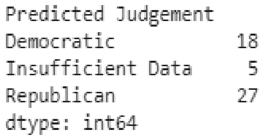


Figure 17. Location based analysis

Results as per location:



```
Predicted Judgement
Democratic          18
Insufficient Data    5
Republican          27
dtype: int64
```

By looking at the results, it seems as per current sentiment, the Republicans have a 27-18 lead. However, the opinion of the 5 states whose sentiment was not clear due to insufficient data.

This analysis can help the Government get an overview of support getting from a particular state. For the states, where the number of supporters are less or have more negative comments, the party can decide to increase the number of campaigns or change campaign strategy which could help them make more positive impact on people resulting in a change in their views towards the Government.

Advantages:

- Performs better as it not only considers the number of tweets but also combines it with the sentiments – both Bing lu and NRC
- Since we considered only one tweet per individual for our analysis, chances of creating a false bias are reduced
- POS tagging and Lemmatizing is used which leads to better accuracy and tweet context.

Disadvantages:

- Cannot correctly classify sarcastic tweets
- Cannot correctly classify sentiments of tweets where both candidate names are present as complex sentence structure

Future Scope:

- Converting this analysis into sentiment score calculator by implementing various Machine Learning Models
- Applying this polarity detection to all kinds of elections and polling
- Applying this analysis to various domains like marketing, transportation, science, journalism, etc to gauge real time sentiments of consumers/people regarding various products/services.

Role of Team Members in the project:

Data Collection	Aditya Gajula	Viharika Bharti	Ritu Gangwal
Database (MySQL)	Aditya Gajula	Viharika Bharti	
Data Pre-processing		Viharika Bharti	Ritu Gangwal
Sentiment analysis (Lexicon)	Aditya Gajula	Viharika Bharti	Ritu Gangwal
Sentiment Analysis (Location)	Aditya Gajula		Ritu Gangwal
Project Report	Aditya Gajula	Viharika Bharti	Ritu Gangwal

References:

[1] Boguslavsky, I. (2017). Semantic Descriptions for a Text Understanding System. In Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference “Dialogue”(2017) (pp. 14-28).

[2] Pak, A., & Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In LREc (Vol. 10, No. 2010).

[3] Scott, J. (2011). Social network analysis: developments, advances, and prospects. Social network analysis and mining, 1(1), 21-26.

[4] Statista, 2017, **https://www.statista.com/statistics/282087/number-ofmonthly-active-twitter-users/**

[5] Wang, H., Can, D., Kazemzadeh, A., Bar, F., & Narayanan, S. (2012, July). A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In Proceedings of the ACL 2012 System Demonstrations (pp. 115-120). Association for Computational Linguistics.

[6] TextBlob, 2017, **https://textblob.readthedocs.io/en/dev/**

[7] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2(1–2), 1- 135.

[8] Dos Santos, C. N., & Gatti, M. (2014, August). Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts

[9] List of Slangs: www.noslang.com/dictionary/

[10] List of emoticons: http://en.wikipedia.org/wiki/List_of_emoticons and
Negations: www.grammarly.com/handbook/sentences/negatives/