BAIT 508 Social Media Analytics Project

JULY 4

MBAN, Sauder School of Business
University of British Columbia
Authored by: Tanishqa Agarwal, Chinmay Jain, Darian Ghorbanin

Individual Contributions

Each of us coded our version of the different parts of the project. We then got together to discuss and incorporated the best versions of each code to get to the best results.

Tanishqa Agarwal

Student Number – 24215832

Email Id: tanishqaagarwal360@gmail.com

Section: BA2

Role: Code for [A], [B], [C], Insights around analysis and Report Formulation

Chinmay Jain

Student Number – 54565486

Email Id: chinmayjain132@gmail.com

Section: BA2

Role: Code for [A], [B], [C], [D] and Insights around analysis

Darian Ghorbanian

Student Number – 79079604

Email Id: darian.ghorbanian@gmail.com

Section: BA2

Role: Code for [A], [B], [C], [D] and Insights around analysis

Project Overview:

People actively express their opinions in social media platforms such as Twitter, Facebook, Instagram, WeChat, TikTok, etc. As aspiring analytics experts, we want to take this opportunity to use our Python skills to conduct social media analysis.

The platform we have chosen for our analysis is Twitter. Twitter has a 229 million active user base across the globe which provides for a lucrative ground to collect ample data from. Every second, on average, around 6,000 tweets are tweeted on Twitter, corresponding to over 350,000 tweets sent per minute, 500 million tweets per day and around 200 billion tweets per year!

Climate change is one of the most discussed topics of the decade. It is not just an environmental crisis anymore but a social issue affecting all sections of society. Some people think it is a big threat to the planet whereas some people think, it is a hoax. In recent years there has been a lot more awareness on this topic with prominent global leaders/celebrities/organizations taking the fore front on addressing the same via their social media platforms.

Many people take to social media to express their support/opinions on such topics when they are especially trending, but what will be interesting to note is the general perception, activity and sentiment around a social issue such as climate change on any given 'normal' day. The aim of this project is to analyze how people's perceptions towards climate change are using sentiment analysis on Twitter data.

Python Libraries Installed

Below code snippet highlights all libraries installed for the purpose of this project –

```
!pip install tweepy
!pip install nltk
!pip install pillow
!pip install wordcloud
!pip install matplotlib
!pip install textblob
%matplotlib inline
!pip install pandas
!pip install numpy
```

```
import string
import re
from PIL import Image
from collections import Counter
import warnings
import pandas as pd
import numpy as np
from textblob import TextBlob
import matplotlib.pyplot as plt
import pickle
import json
from pprint import pprint
from wordcloud import WordCloud
from TwitterCollector import TwitterCollector
from datetime import datetime
import nltk
nltk.download('punkt')
nltk.download('stopwords')
```

A. Keyword Selection and Data Collection

1. Apply for a Twitter Developer's account

Once we got the account we created a dummy app, and from that app we got the necessary Keys & Tokens which we need for the API.

2. Connecting with Twitter API and extracting the data

After the connection is established, we can query the Twitter API and ask for the data we want.

```
# Enter bearer token
bearer_token = r"bearer token"

# initialize a TwitterCollector instance
tc = TwitterCollector(bearer_token = bearer_token)
```

We chose 'Climate Change' as our main keyword to pull 10k recent tweets from Twitter. We made sure to filter out any retweets and filtered for tweets only in English. The extracted data was then saved into a json file.

```
query1 = '"Climate Change" -is:retweet lang:en'

recent_tweets = tc.fetch_recent_tweets(query = query1
                                       , tweets_cnt = 10000
                                       , save_result = True
                                       , save_dir = 'sample_data'
                                       , file_name ='ClimateFinal.json'
```

Below is the json file with 10k tweets on climate change

3. Getting the list of unique author IDs

We open the json file created above to do our analysis going forward using the code below -

```
import json
import pprint

with open('ClimateFinal.json') as json_file:
    json_cc_data=json.load(json_file)
```

Below code helps us get the list of unique author IDs

```
auth_id=[]
for tweet in json_cc_data['tweets']:
        if tweet['author_id'] not in auth_id:
            auth_id.append(tweet['author_id'])
len(auth_id)
```

This gives us a list of **8258** unique author ids from the 10k tweets collected

4. **Collecting author information for the collected author IDs**

Because there is limitation in fetching huge amounts of data at a time, we are fetching author information for 300 authors at a time. Snippet below –

```
ll=0
import time
if len(author)+ll+300 <= len(auth_id):
    l=len(author)+ll+300
else:
    l=len(auth_id)

for i in range(ll+len(author),l):
    if i!= 4459:
        author.append(tc.fetch_author_info(auth_id[i]))
```

We then take only the relevant columns from the author data and create a pandas data frame for ease of analysis and then finally save it into a csv file.

It is worth mentioning as some of the authors were deleted, when we encounter the error that object not existed, we found the relevant i, and changed it in the above if statement. Moreover, our data is collected in 4 different times, and we have 4 different csv files. We finally merged all of them. The relevant notebook for merging is also attached.

```
import pandas as pd
author_df=pd.DataFrame({'author_id':a_id,
                        'name':name,
                        'username':username,
                        'location':location,
                        'created_at':created_at,
                        'description':description,
                        'followers_count':followers_count,
                        'following_count':following_count,
                        'tweet_count':tweet_count,
                        'listed_count':listed_count,
                        'verified':verified,
                        'description':description})
```

Below is the final csv file that contains all relevant author information

Author_Info.csv

B. Preliminary Analysis

1. 10 Most Popular words with and without stop words

- Without stopwords:

  Below are the top 10 most popular words from the twitter data set extracted (recent_tweets)

```
[('the', 11405),
 ('climate', 9846),
 ('change', 8887),
 ('to', 7789),
 ('and', 6063),
 ('of', 5441),
 ('is', 5262),
 ('a', 4683),
 ('in', 3690),
 ('for', 2647)]
```

- With stopwords:

  We used the list below to add further stopwords into the already existing stopword list imported from nltk library

```
words = ['climate','change','Climate', 'The', '&', '&amp',
         '&amp;','change.', 'change,', 'Change',
         'like', 'We', 'You', 'And',
         'This', 'us', 'know', 'would', 'make', 'think', 'If', 'It',
         'get', 'They', 'change?', 'What', 'going',"it's", 'want',
         'even', 't', 'amp', 'https:', 'co', 'https']

for w in words:
    if w not in stopwords and len(w) > 1:
        stopwords.append(w)

print(len(words))
print(stopwords)
```

Below are the top 10 most popular words from the twitter data set extracted (recent_tweets)

```
[('climate', 9846),
 ('change', 8887),
 ('people', 885),
 ('global', 547),
 ('world', 469),
 ('one', 455),
 ('hurricane', 433),
 ('real', 396),
 ('need', 370),
 ('years', 367)]
```

2. **10 Most Popular word hashtags**

Below are the 10 most popular hashtags from the twitter data extracted(recent_tweets)

```
[('#climatechange', 206),
 ('#climate', 156),
 ('#freya', 55),
 ('#freyathewalrus', 55),
 ('#climatecrisis', 51),
 ('#hurricaneian', 45),
 ('#climateemergency', 44),
 ('#climateaction', 42),
 ('#cop27', 32),
 ('#climatescam', 27)]
```

Insight: It is interesting to note that Hurricane Ian originated on 19th Sept and 15 days later, is still a trending hashtag within climate change.

Also, interesting are the hashtags surrounding Freya the Walrus**.** She was euthanized mid-August, however there is still considerable chatter around the topic within climate change.

3. **10 Most frequently mentioned usernames**

   Below are the 10 most frequently mentioned usernames from the twitter data extracted(recent_tweets)

```
[('@foxnews', 88),
 ('@tomfitton', 79),
 ('@elonmusk', 69),
 ('@govrondesantis', 62),
 ('@stevenbeschloss', 59),
 ('@timrunshismouth', 53),
 ('@potus', 53),
 ('@petersweden7', 53),
 ('@jojofromjerz', 48),
 ('@totalenergies', 48)]
```

   Insight: The mix includes 1 news channel, 8 influential twitter personalities, and 1 company that produces & markets energies on a global scale.

4. **3 Most common sources of tweet**

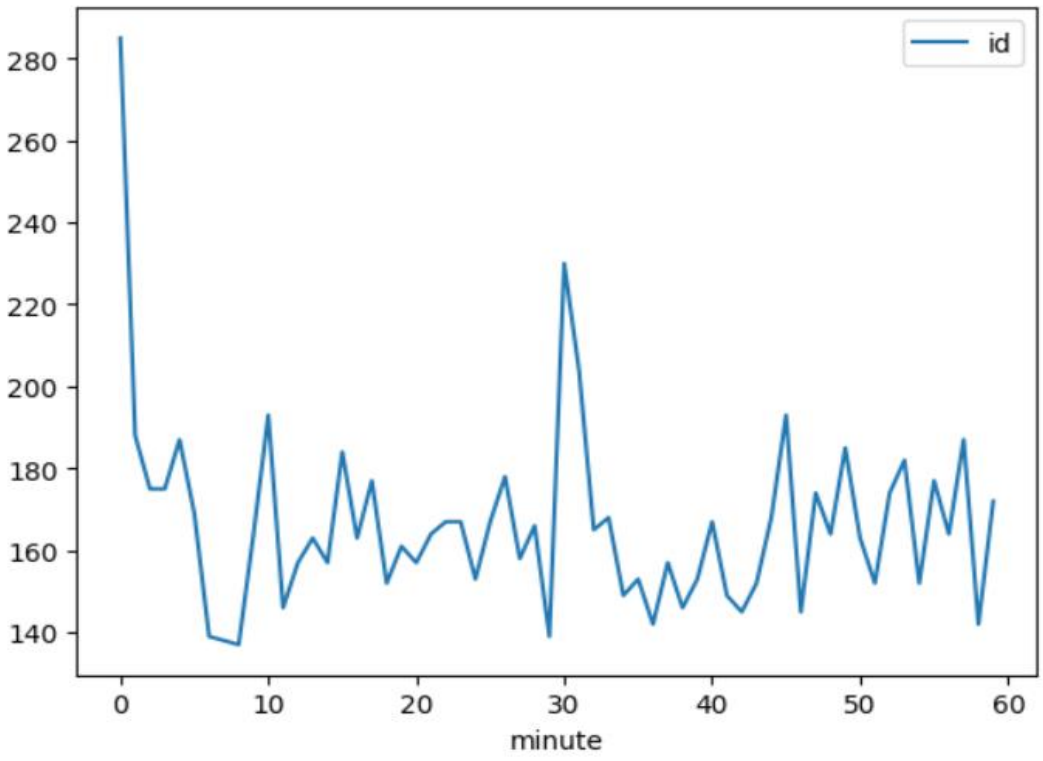
   Below are the 3 most common sources of tweet from the twitter data set extracted(recent_tweets)

```
[('Twitter Web App', 3909),
 ('Twitter for iPhone', 2530),
 ('Twitter for Android', 1966)]
```

   Insight: The general assumption is that people mainly use their phones to access social media especially when it comes to quick posts such as twitter, however the twitter web app showing up as the top source of tweet makes for an interesting insight.

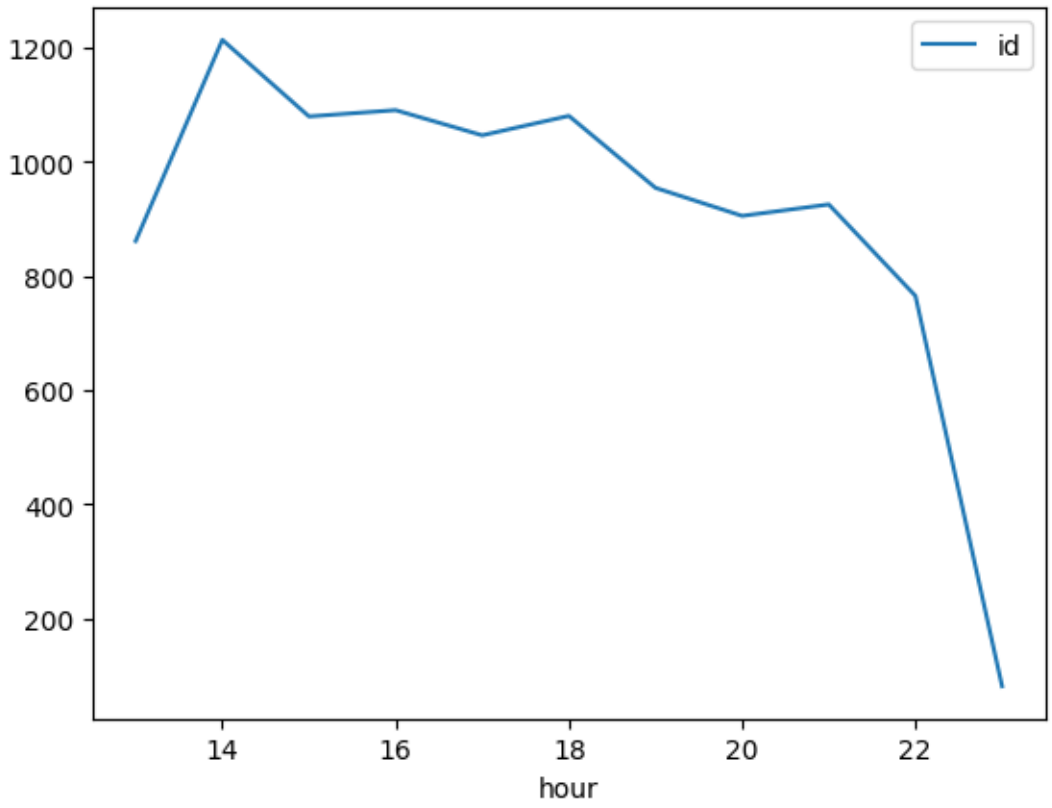
5. **Time trend of tweet counts per minute of the hour**

   The below trend line highlights the number of tweets posted at each minute of the hour across all hours in a day



Insight: There are prominent spikes at the 0th minute, i.e. (11:00 AM, 12:00 PM, etc.) and the 30th minute, i.e. (11:30 AM, 12:30 PM, etc.)

The below trend highlights the number of tweets posted at each hour of the day



Insight: We can see that people mostly tweet in the afternoon and as the day progresses the no. of tweets reduces. Of the 10,000 tweets we collected 12% of the tweets were made at 2 pm

6. 3 Most influential tweets

A tweet's influence score is the sum of "quote_count", "reply_count", "retweet_count", "like_count"

```
7314    So I was the only WH pool print reporter in th...
9106    The same people telling you to stop climate ch...
7040    .@KwasiKwarteng has gall to say he has sound p...
```

7. 3 Most vocal authors

We chose the authors with the maximum tweets within data collected (recent_tweets) to come up with the top 3 most vocal authors

```
Peter Paul              55
Seeking Truth&Facts     72
leave world better      31
```

8. **3 Most Influential authors**

   A user's influence score is the sum of "followers_count", "following_count", "listed_count". "tweet_count" in the author data

```
5756              CNN
2412    The Economist
4489          Reuters
```

   Insight: The top 3 most influential authors from our dataset collected are all prominent news channels

C. Word Cloud

We have used the following functions to enhance our word cloud:

- Mask, to change shape of default word cloud to custom shape as per image loaded
- 'Scale' of 3, which makes the computation faster for larger word cloud images
- 'Colormap' is set to 'Blues' to better suit to our image shape and selected keyword for the analysis
- Collocations are set to False to give a fuller word cloud
- Max word is set to 125 to avoid clutter in the cloud
- We replaced certain words/punctuations from our data and then added certain redundant words to the stopword list to create a clean word cloud with relevant words

Code Snippet –

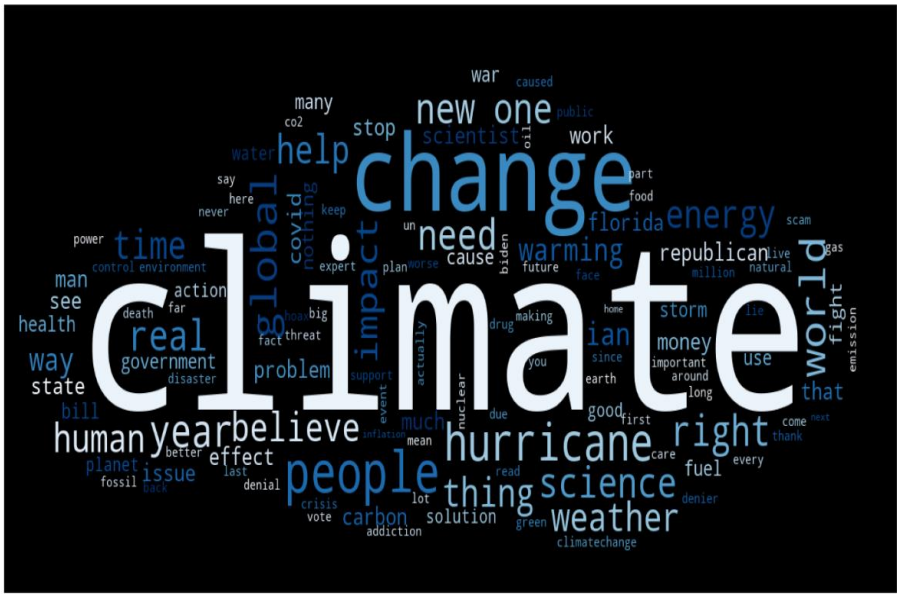
```
wordcloud1 = wordcloud1.replace("don’t", "")
wordcloud1 = wordcloud1.replace("i’m", "")
wordcloud1 = wordcloud1.replace("!", "")
wordcloud1 = wordcloud1.replace("\'", "")
wordcloud1 = wordcloud1.replace("didn’t", "")
wordcloud1 = wordcloud1.replace(" let ", "")
wordcloud1 = wordcloud1.replace(" go ", "")
wordcloud1 = wordcloud1.replace(" etc ", "")
wordcloud1 = wordcloud1.replace(" well ", "")
wordcloud1 = wordcloud1.replace(" still ", "")
wordcloud1 = wordcloud1.replace(" today ", "")
wordcloud1 = wordcloud1.replace(" day ", "")
wordcloud1 = wordcloud1.replace(" must ", "")
wordcloud1 = wordcloud1.replace(":", " ")
wordcloud1 = wordcloud1.replace(" od ", " ")
wordcloud1 = wordcloud1.replace('"', '')
wordcloud1 = wordcloud1.replace('"climate', 'climate')
wordcloud1 = wordcloud1.replace(" im ", " ")
wordcloud1 = wordcloud1.replace(" re ", "")
wordcloud1 = wordcloud1.replace(" s ", "")
wordcloud1 = wordcloud1.replace(" S ", "")
wordcloud1 = wordcloud1.replace("change”", "change")
wordcloud1 = wordcloud1.replace(" we ", "")
```

```
cloud_image = r'/Users/Tanishqa Agarwal/Downloads/cloud.jpg'

image = np.array(Image.open(cloud_image))
word_cloud = WordCloud(max_words = 125,
                       max_font_size = 100,
                       collocations = False,
                       scale =3,
                       mask = image,
                       colormap = 'Blues',
                       stopwords=['re', 'S', 't', 'we', 'could', 'also']).generate(wordcloud1)

# Display the generated image:
plt.figure(figsize=(20,10))
plt.imshow(word_cloud, interpolation="bilinear")
plt.axis("off")
plt.savefig('my_word_cloud.png')
plt.savefig('my_word_cloud.pdf')
plt.show()
```

Below word cloud highlights the most significant 125 words from our twitter data source



D. Sentiment Analysis

1. Average Polarity and Subjectivity scores

We used the TextBlob python library to run our sentiment analysis on the data collected. TextBlob is a python library for Natural Language Processing (NLP). TextBlob actively uses Natural Language Tool Kit (NLTK) to achieve its tasks.

Polarity lies between [-1,1]. A polarity of -1 indicates a negative sentiment whereas a +1 indicates a positive sentiment.

Subjectivity lies between [0,1]. Subjectivity quantifies the amount of personal opinion and factual information contained in the text.

Code Snippet –

```
twitter_data['polarity'] = None
twitter_data['subjectivity'] = None
for i in range(0,len(twitter_data)):
    twitter_data['polarity'][i] = TextBlob(twitter_data['text'][i]).sentiment.polarity
    twitter_data['subjectivity'][i] = TextBlob(twitter_data['text'][i]).sentiment.subjectivity
```

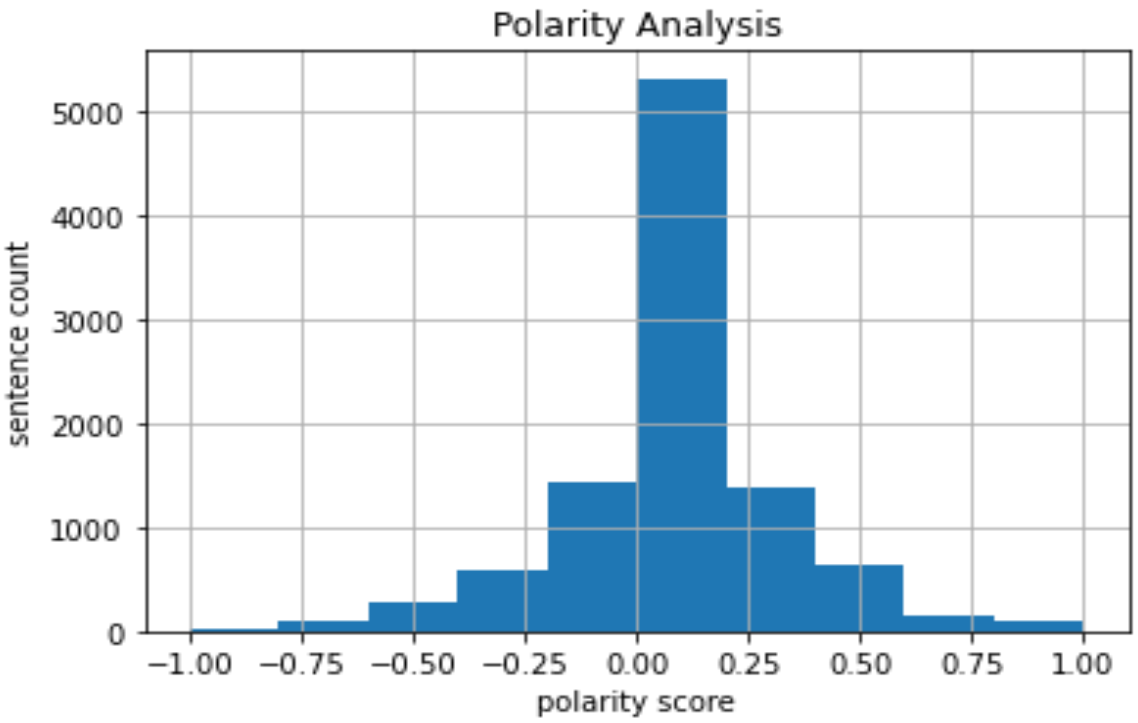
Below is the avg. polarity and subjectivity of our dataset(recent_tweets) respectively-

```
Average Polarity score across the data is: 0.07
Average Subjectivity score across the data is: 0.39
```

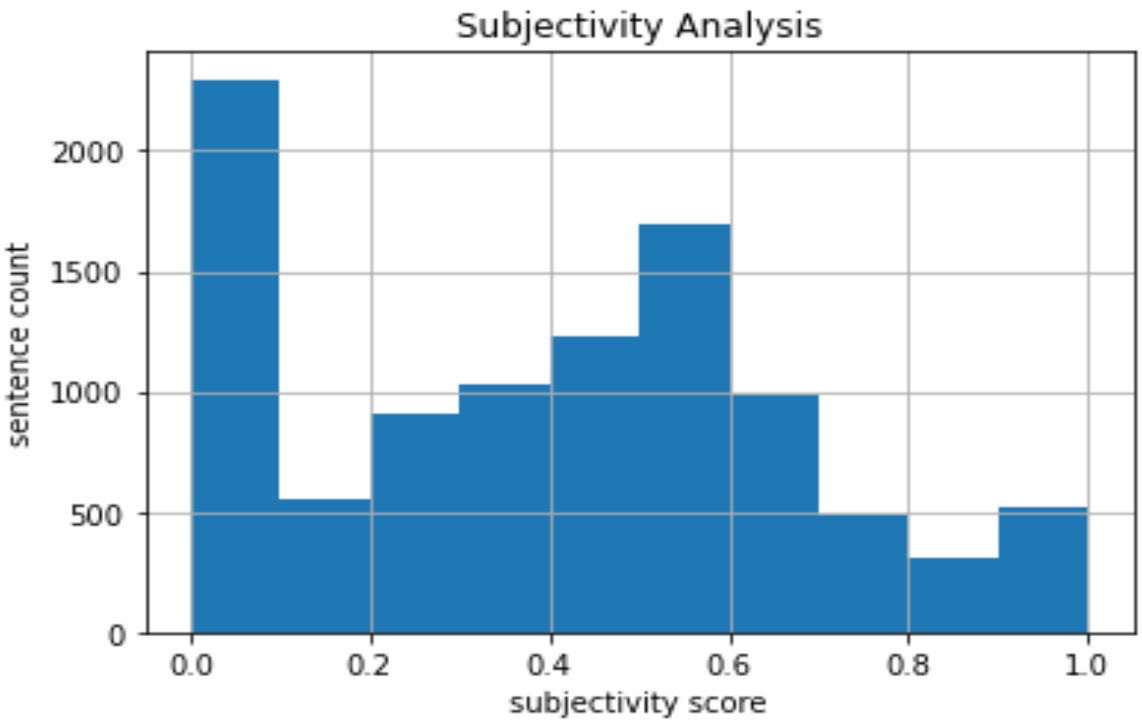
Insight: The avg. polarity score of 0.07 indicates a somewhat neutral sentiment around climate change within the data set we collected, i.e., neither positive nor negative.
The avg. subjectivity score of 0.39 indicates that the data contains relatively higher personal opinion as compared to factual information.

2. Polarity and Subjectivity Score Distributions

Polarity Scores Histogram –



Subjectivity Scores Histogram –



Insight: The polarity score distribution ties back with our mean score indicating that the number isn't skewed, and the general sentiment is neutral around climate change
The subjectivity score distribution shows a peak between 0.0 & 0.1 indicating that ~2.2 k tweets (i.e., 25% of the tweets collected) are purely factual. However due to a wider distribution of tweets between 0.4 & 0.6, this insight gets masked in the mean subjectivity score despite accounting for 25% of the data.

3. Tweets with Most Positive & Negative Polarity

- Top 3 tweets with Highest Polarity

```
Polarity Score:  1.0
 @BitcoinLovesAll @Reuters Coming from a bitcoin &amp;amp; climate change denier clown spamming garbage: priceless.
Polarity Score:  1.0
 @mattgaetz Dear FL, your elected reps have done nothing to prepare for climate change and are now asking for a handout. Whi
le we are very happy to help our fellow US citizens during a crisis, please address the long-term sustainability of inhabiti
ng a sinking ship.
Polarity Score:  1.0
 @WSandsin @GovRonDeSantis @CaseyDeSantis You need to focus on something besides "climate change" (Democrats BS) and the in
s. companies are going backrupt is why they are gone. Ron is the best Governor in the United States.
```

Insight: Reading the tweets, we can infer that the authors for their respective tweets are actually NOT happy about climate change, however the tweet uses certain very positive words that marks these as tweets with polarity 1

- Top 3 tweets with Highest Polarity

```
Polarity Score:  -1.0
 @KTFministry Notice that all the discussion is about climate change &amp;amp; man's use of oil but SIN IS never mentioned. The
Bible shows us that evil was on the earth and God destroyed Sodom &amp;amp; Gomorrah. Sulfur balls are still found there but the
re is no life even to this day in that part of world.
Polarity Score:  -1.0
 You've all been brainwashed into thinking climate change is terrifying and nuclear war with Russia isn't
Polarity Score:  -1.0
 For the Brainwashed Morons fretting about Non-existent Climate Change. Read this and shut the hell up!
```

Insight: Reading the tweets, we can infer that the authors are not very happy and expressing their negative sentiment about the topic is evident from both the polarity score and the text of the tweet as well.

E. Insights

We did some additional analysis on the data that was collected to better understand the data and draw some interesting insights from it.

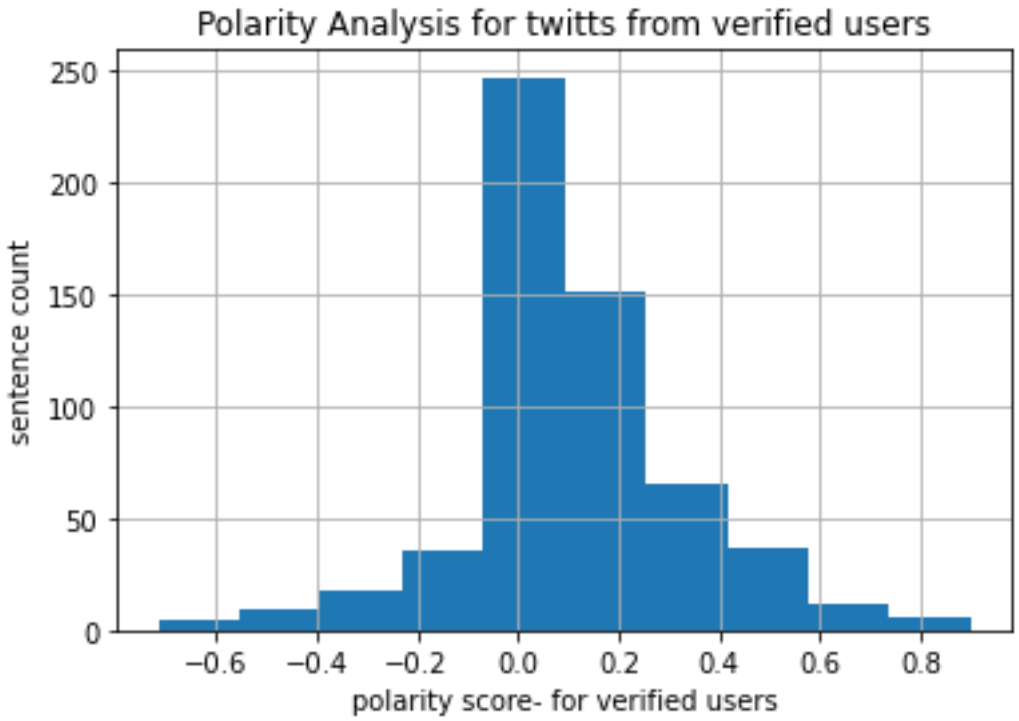
Verified & Unverified Authors

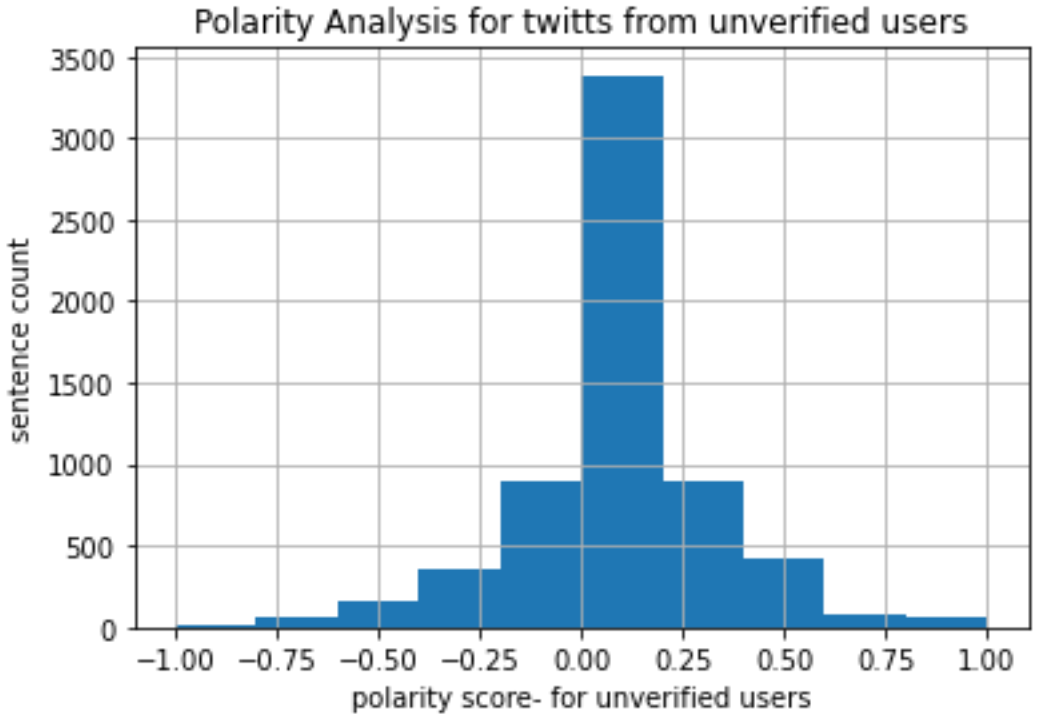
93.6% of the authors, i.e., 7729 of the authors are unverified (without a blue tick on twitter) and 6.4%, i.e., 528 are verified authors (with a blue tick on twitter). What will be interesting to note are the polarity and subjectivity scores associated with each group.

verified	polarity	subjectivity
False	0.073826	0.394526
True	0.105565	0.387753

The avg. subjectivity score is pretty similar for both verified and unverified users, however the avg. polarity score sees some difference, indicating verified users in the dataset tend to be slightly more mindful/positive about their tweets compared to unverified users.

This can be further corroborated using the polarity score histograms for verified and unverified users as below –





During the report above, we added each part of insights below each of the analysis which was important. Moreover, based on those insights and further analysis, we can add the following insights as well:

- Tweets done via web app have on average higher sentiment. This may show that people do differently when they use their laptops vs. when they are with their phones.
- Although the top vocal authors are publications, the highest influential tweets are done by individuals. It may seem that individuals can have affect more in comparison to publications.
- Two authors who had made the top 10 tweets regarding influence score have deleted their accounts two days after. So, we can find that there are some people that make accounts and make tweets which are influential, but then delete their accounts. They can be some special people organized by some politicians or other people.
- We can see some subtle difference for sentiment analysis of verified and unverified analysis, but there is no significant difference in both means for sentiments and also plots.