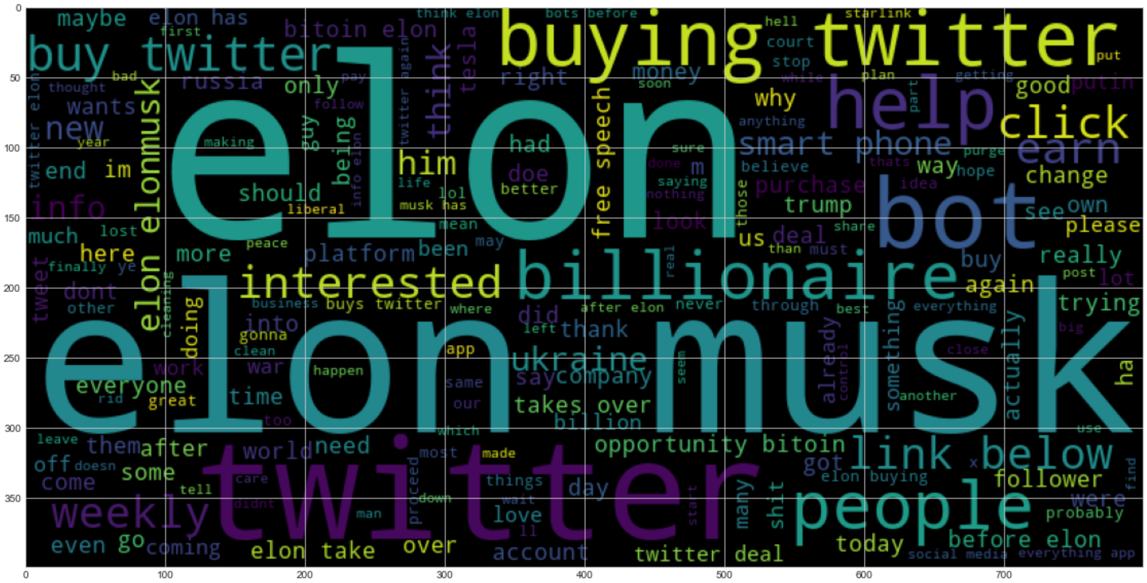


Social Media Analysis

BAIT 508 | October 5, 2022

Part A

Project Description

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 by using our Python skills to conduct social media analysis.

For the purposes of this project, we used twitter data. Below are a few details regarding the overall analysis and the keywords considered.

- We considered elon as our keyword, avoiding tweets in languages other than English and retweets. Retweets can really skew our data, considering the same text repeatedly.

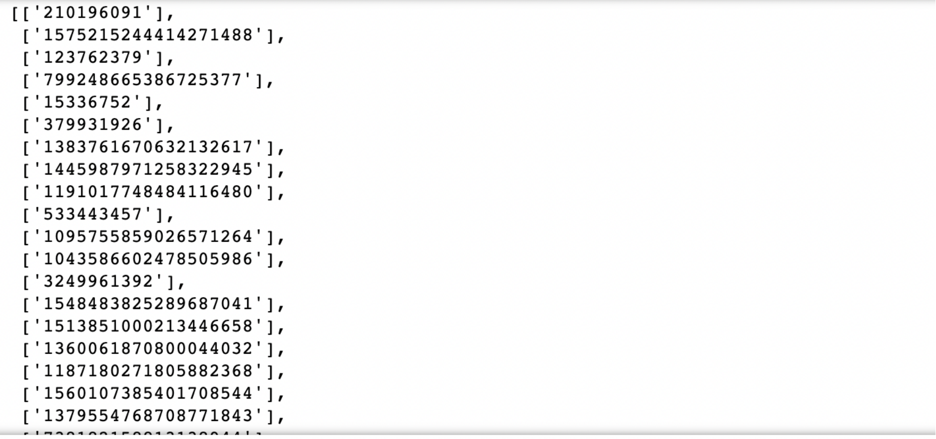
- We used twitter API to get ~10,000 based on the keyword, as Elon Musk put his bid forward for buy twitter.

Preliminary Analysis

- Keyword: elon

- Based on the data collected, we get ~8000 distinct authors. We further collect author info using TwitterCollector.py

Result:



```
[['210196091'],
 ['1575215244414271488'],
 ['123762379'],
 ['799248665386725377'],
 ['15336752'],
 ['379931926'],
 ['1383761670632132617'],
 ['1445987971258322945'],
 ['1191017748484116480'],
 ['533443457'],
 ['1095755859026571264'],
 ['1043586602478505986'],
 ['3249961392'],
 ['1548483825289687041'],
 ['1513851000213446658'],
 ['1360061870800044032'],
 ['1187180271805882368'],
 ['1560107385401708544'],
 ['1379554768708771843'],
```

- Author id information are extracted in ques 7 and 8 in part B

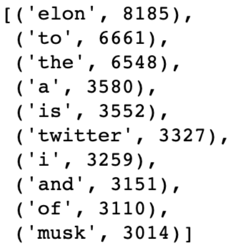
Part B

- For Ques 1-4, we use an empty list and append the required words in the same.
    - In Ques 1, we use the for loop and split command to store all the words in the words_goat list.
    - We continue to use words_goat to extract specific word types based on the requirements of the rest of the questions

Ques 1. What are the ten most popular words with and without stop words?

- With Stopwords
    - elon is most popular word in our tweets.
    - In the below code, we have included stopwords and hence we see words like 'to' and 'a' in our results.
    - It is interesting to note that despite the stopwords, elon, musk and twitter are heavily present in our sample. This tells us that majority of the tweets could be regrading elon's twitter buyout bid, but we are yet to confirm that assumption.

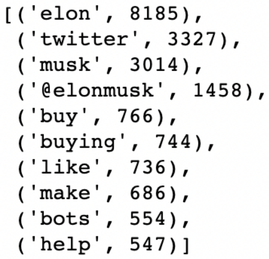
Output:



```
[('elon', 8185),
 ('to', 6661),
 ('the', 6548),
 ('a', 3580),
 ('is', 3552),
 ('twitter', 3327),
 ('i', 3259),
 ('and', 3151),
 ('of', 3110),
 ('musk', 3014)]
```

- Without Stopwords
    - We use the stopwords pickle file to remove majority of stopwords and get a better insight regarding the tweets and the general topics our sample was discussing.
    - We observe that elon, musk, buy, buying and twitter are one of the most of tweets, further suggesting that most of the tweets could be about elon musk's bid to buy twitter.

Output:

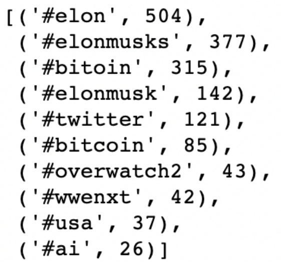


```
[('elon', 8185),
 ('twitter', 3327),
 ('musk', 3014),
 ('@elonmusk', 1458),
 ('buy', 766),
 ('buying', 744),
 ('like', 736),
 ('make', 686),
 ('bots', 554),
 ('help', 547)]
```

Ques 2. What are the ten most popular hashtags (#hashtag)?

While #elonmuks and #elon are the most used hastags, the other hashtags given us an insight on the general topic of discussion, which involves twitter and bitcoin.

Output:

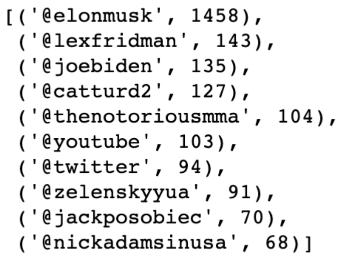


```
[('#elon', 504),
 ('#elonmusks', 377),
 ('#bitoin', 315),
 ('#elonmusk', 142),
 ('#twitter', 121),
 ('#bitcoin', 85),
 ('#overwatch2', 43),
 ('#wwenxt', 42),
 ('#usa', 37),
 ('#ai', 26)]
```

Ques 3. What are the ten most frequently mentioned usernames (@username)?

While elon musk is the most mentioned person on the tweets, we notice that lex fridman, who is an AI researcher and has a Tech podcast. This tells us that apart from the twitter conversation, a lot of tweets also talk about AI and the deep mind (elon's venture towards AI). Joe Biden has also been mentioned multiple times.

Output:

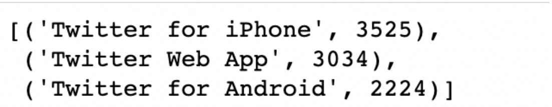


```
[('@elonmusk', 1458),
 ('@lexfridman', 143),
 ('@joebiden', 135),
 ('@catturd2', 127),
 ('@thenotoriousmma', 104),
 ('@youtube', 103),
 ('@twitter', 94),
 ('@zelenskyyua', 91),
 ('@jackposobiec', 70),
 ('@nickadamsinusa', 68)]
```

Ques 4 Which are the three most common sources of the tweets?

- Most of the authors in our sample used twitter for iphone and the least amount of people used twitter for android

Output:

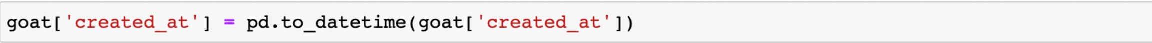


```
[('Twitter for iPhone', 3525),
 ('Twitter Web App', 3034),
 ('Twitter for Android', 2224)]
```

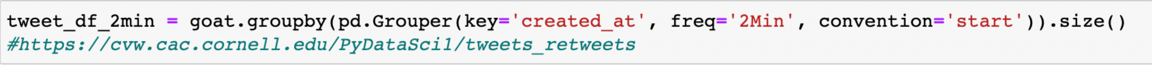
Ques 5. Create a line chart to show the time trend of tweet counts (number of tweets in a day (or an hour or a minute) depending on the collected data)

We used the grouper function in the pandas library to define the frequency and convention of the dataset. Size function tells us the number of tweets.

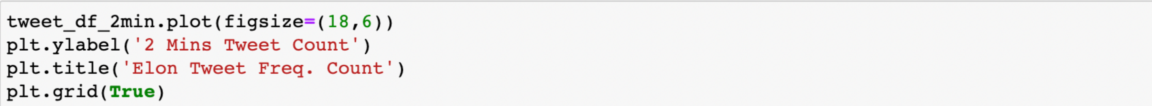
Code:



```
goat['created_at'] = pd.to_datetime(goat['created_at'])
```

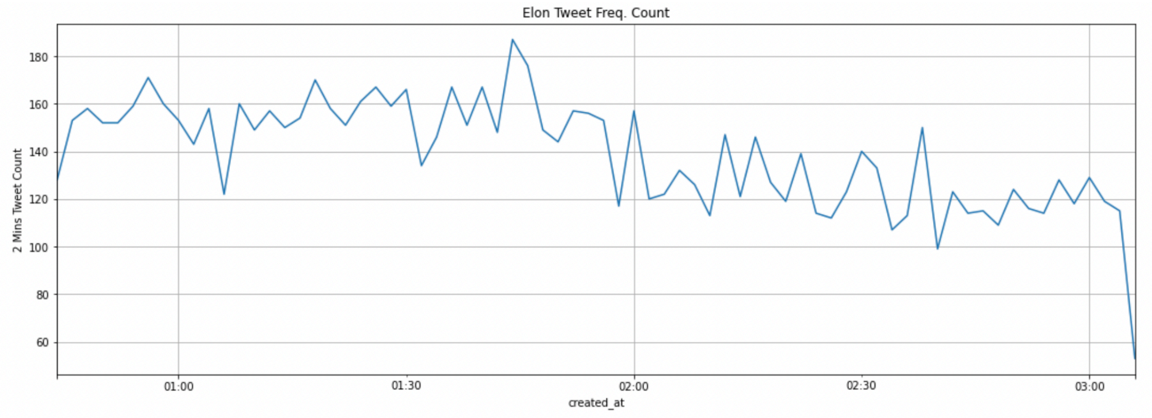


```
tweet_df_2min = goat.groupby(pd.Grouper(key='created_at', freq='2Min', convention='start')).size()
#https://cvw.cac.cornell.edu/PyDataSci1/tweets_retweets
```



```
tweet_df_2min.plot(figsize=(18,6))
plt.ylabel('2 Mins Tweet Count')
plt.title('Elon Tweet Freq. Count')
plt.grid(True)
```

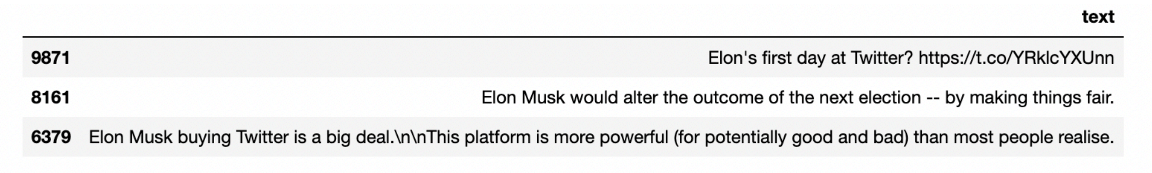
Output:



The above chart shows us the number of tweets every 2 minutes vs the tweet date. Given we only considered 10,000 tweets for our analysis, the number of tweets are only from ~4 hours.

Ques 6. Which are the three most influential tweets?

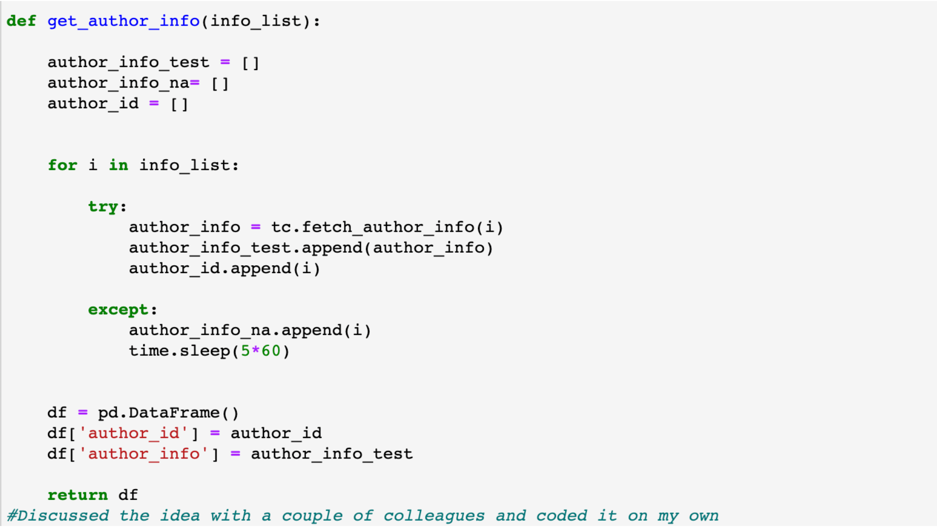
Below are the most influential tweets based on the influence score (A tweet's influence score is the sum of "quote_count", "reply_count", "retweet_count", "like_count")



	text
9871	Elon's first day at Twitter? https://t.co/YRklcYXUnn
8161	Elon Musk would alter the outcome of the next election -- by making things fair.
6379	Elon Musk buying Twitter is a big deal.\n\nThis platform is more powerful (for potentially good and bad) than most people realise.

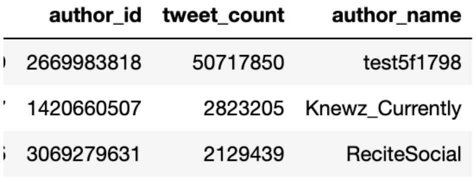
Ques 7. Who are the three most vocal authors on the keyword? In other words, who are the most frequently tweeting authors in the tweet data?

Below are the most frequently tweeting authors. We used Twitter collector to extract author info. We divided the data into multiple sets of 500 author ids and created a function to extract author info. The Set was divided into multiple parts due to the "too many requests" error.



```
def get_author_info(info_list):

    author_info_test = []
    author_info_na= []
    author_id = []


    for i in info_list:

        try:
            author_info = tc.fetch_author_info(i)
            author_info_test.append(author_info)
            author_id.append(i)

        except:
            author_info_na.append(i)
            time.sleep(5*60)


    df = pd.DataFrame()
    df['author_id'] = author_id
    df['author_info'] = author_info_test

    return df
#Discussed the idea with a couple of colleagues and coded it on my own
```

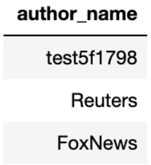
Based on the merged df, we used sort values to get the users with the max tweet count. Below are the most frequently tweeting authors in our sample.



author_id	tweet_count	author_name
2669983818	50717850	test5f1798
1420660507	2823205	Knewz_Currently
3069279631	2129439	ReciteSocial

Ques 8. Who are the three most influential authors?

Below are the most influential authors based on the influence score (A tweet's influence score is the sum of "quote_count", "reply_count", "retweet_count", "like_count")

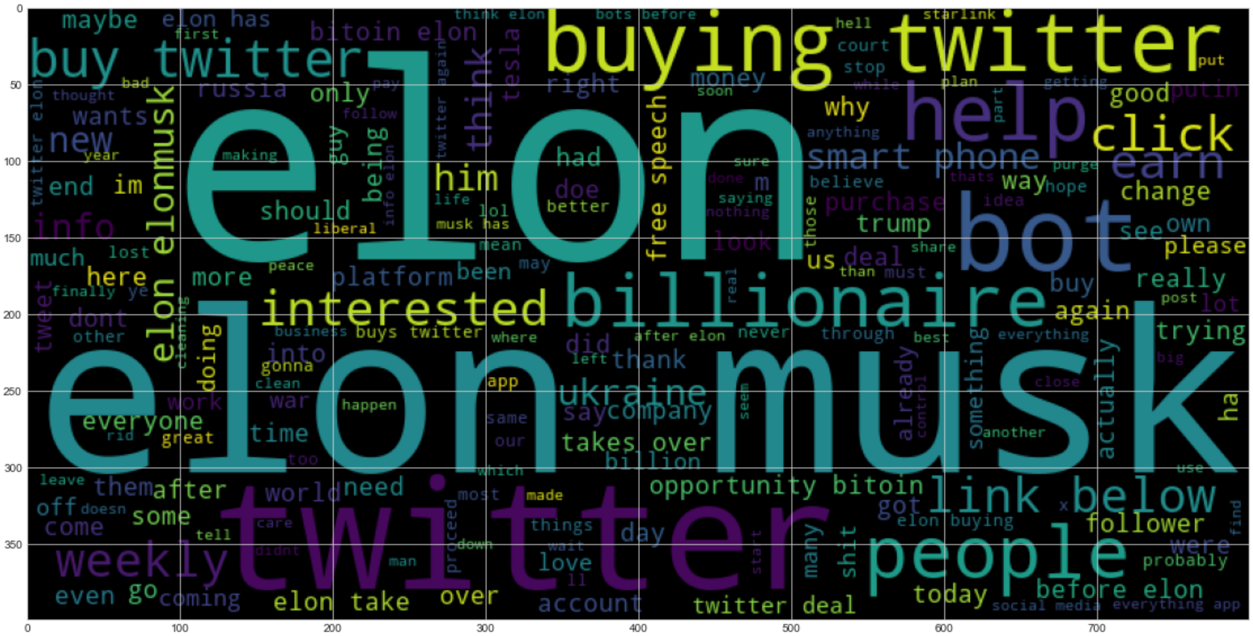


author_name
test5f1798
Reuters
FoxNews

Part C - Word Cloud

Word cloud helps us look at the general discussion topic within our sample.

- Most of the sample is talking about elon musk’s bid to buy twitter
- A lot of people are also talking about bitcoin and the opportunity it presents for Elon



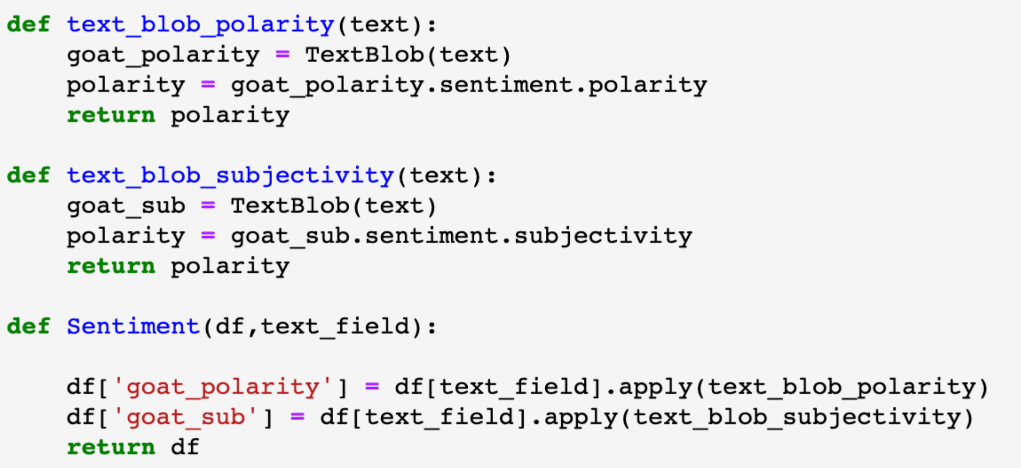
- Elon musk and buying twitter are our top words. Clearly majority of the conversation in our sample around Elon Musk’s bid to by twitter
- We also observe people are commenting about trump, Ukraine and putin in the wordcloud
- Bot is also another highly used keyword in our file, with some people pointing out how another spread of elon’s bid to buy twitter could be due to bot’s spreading information

Part D – Sentiment Analysis

The next step in our analysis is to understand the overall sentiment of our sample. We used textblob package in order to analyze the tweets. We use our twitter data to understand the polarity and subjectivity of the tweets.

- Polarity: Polarity ranges from -1 to 1 depending upon on positive or negative emotion portrayed in the tweet
- Subjectivity: Subjectivity ranges from 0 to 1, indicating the judgements and opinions

In order to calculate sentiment, we created 3 loops to get polarity subjectivity and finally combine all the fields in the Date Frame



```python
def text_blob_polarity(text):
    goat_polarity = TextBlob(text)
    polarity = goat_polarity.sentiment.polarity
    return polarity

def text_blob_subjectivity(text):
    goat_sub = TextBlob(text)
    polarity = goat_sub.sentiment.subjectivity
    return polarity

def Sentiment(df,text_field):

    df['goat_polarity'] = df[text_field].apply(text_blob_polarity)
    df['goat_sub'] = df[text_field].apply(text_blob_subjectivity)
    return df
```

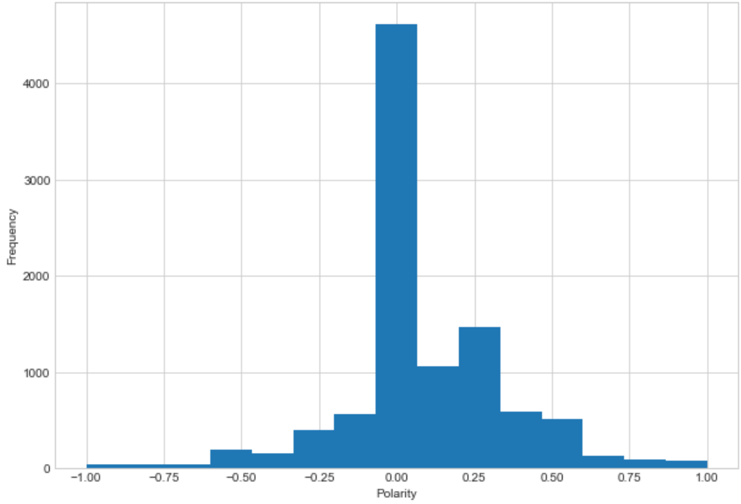
Then we called the above function on our twitter text field (which contains the tweet) to get the sentiment scores. Below we have answered some important questions regarding the sentiment analysis

Ques 1. What are the average polarity and subjectivity scores?

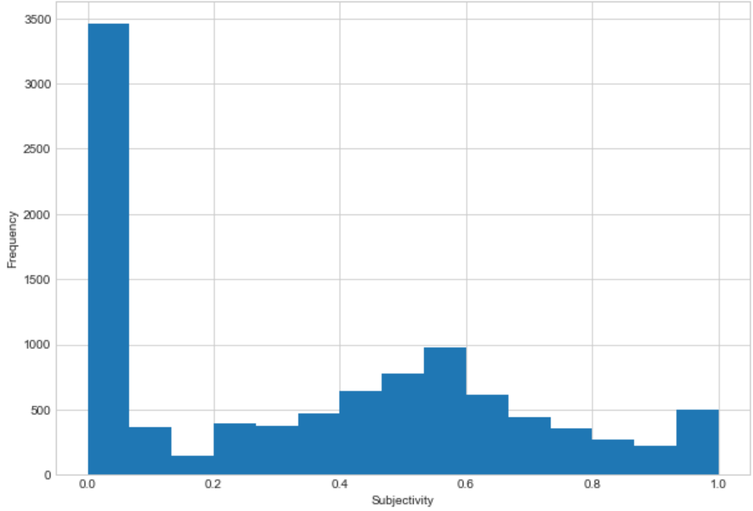
- Average Polarity Score: 0.079
- Average Subjectivity Score: 0.351

Ques 2. Visualize the polarity and subjectivity score distributions using histograms, where X-axis is the score, and Y-axis is the tweet count in the score bin.

- Polarity Histogram
    - Majority of the polarity scores ato and we see more tweets towards the positive side

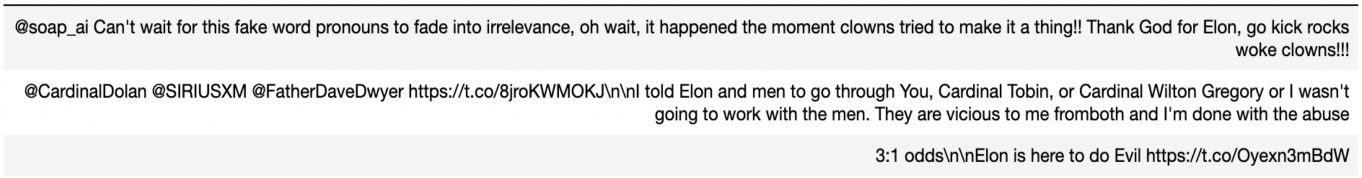


- Subjectivity Histogram
    - Most of the tweets have the subjectivity score of 0. We see the second peak between 0.5 and 0.6.
    - It is interesting to note that ~ 500 tweets have a subjectivity score of 1



Ques 3. Based on the polarity scores, what are the most positive and negative tweets on the keyword? Why is the author happy/angry on the topic? If there are multiple tweets with same sentiment scores, please pick 2-3 tweets among them.

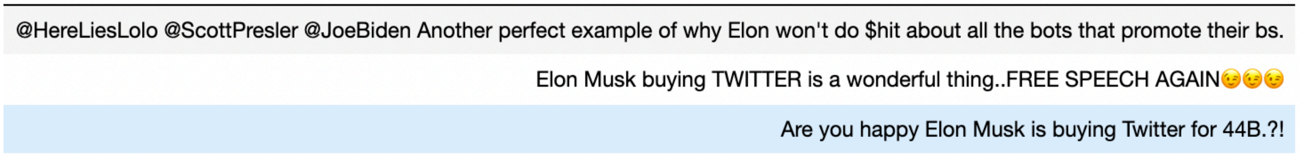
- Most Negative Tweets (based on the sentiment analysis
    - Below are 3 examples of the most negative tweets in our sample



@soap_ai Can't wait for this fake word pronouns to fade into irrelevance, oh wait, it happened the moment clowns tried to make it a thing!! Thank God for Elon, go kick rocks woke clowns!!!
@CardinalDolan @SIRIUSXM @FatherDaveDwyer https://t.co/8jroKWMOKJ\n\nI told Elon and men to go through You, Cardinal Tobin, or Cardinal Wilton Gregory or I wasn't going to work with the men. They are vicious to me fromboth and I'm done with the abuse
3:1 odds\n\nElon is here to do Evil https://t.co/Oyexn3mBdW

- The First tweet is applauding the fact that Elon is buying twitter but still comes in the most negative tweet
- The inability of sentiment analysis to understand sarcasm or positive tweets based on the connotation leads to this tweet being negative

- Most Positive Tweets
    - Below are 3 examples for the most positive tweets



@HereLiesLolo @ScottPresler @JoeBiden Another perfect example of why Elon won't do $hit about all the bots that promote their bs.
Elon Musk buying TWITTER is a wonderful thing..FREE SPEECH AGAIN😉😉😉
Are you happy Elon Musk is buying Twitter for 44B.?!

- The above tweets also signify the inability of the sentiment analysis to correctly classify the tweets
- The second tweet shows that the author is happy about Elon buying twitter
- Tweet number 1 seems very angry but is classified as a positive tweet. Tweet number 3 is a sarcastic but is wrongly classified due to the inability of sentiment analysis to understand sarcasm

Part E – Insights (Part A)

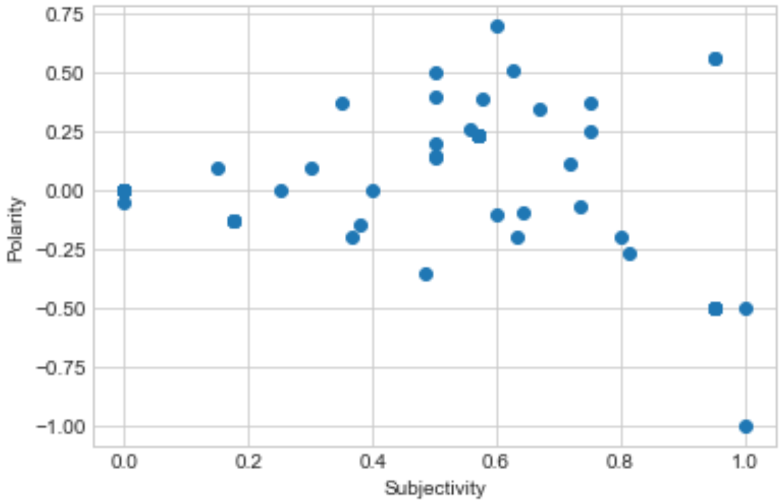
- Elon Musk ‘s original and revived bid to acquire Twitter has helped him create a real buzz among twitter users, as three out of the top ten most popular words and two out of the three most influential tweets are all directly related to this topic.

- With @joebiden and @zelenskyyua being among the ten most frequently mentioned usernames, twitter users seem to associate Elon Musk with the political world as well as influential figures. Twitter users are apt to place Elon Musk on the right of the political spectrum since they tend to bring up well-known right-wing figures such as Donald Trump (word cloud), Jack Posobiec (@jackposobiec) and Nick Adams (@nickadamsinusa) alongside with Elon Musk.

- With way more tweets scoring 0 for both polarity and subjectivity, it suggests most tweets in relation to Elon Musk are considered to be neutral in nature and are not subject to too much subjective opinions of the tweeter users.

- Although Elon Musk is known for his company, Tesla and his innovation and AI capabilities, the most trending topic from the word cloud and happy and unhappy tweets show that most Twitter users post more about his association with Twitter than with other topics of discussion.

- Out of the 10 most used hashtags, 23% of them are related to bitcoin as well. Tweets regrading bitcoin have ~0.11 average polarity and ~0.5 average subjectivity, showing that tweets are well balanced between both factual and opinionated. Most of the tweets are in a positive light



Part E – Part B

Social media analysis presents a lot of opportunities for growing as well as established businesses. With the use to tik-tok, Instagram reels and twitter to promote brands and highlighting the overall use of a product, businesses are improving their overall brand image as well as their profits. But it also presents several challenges as well such as correct allocation of resources, profit against the marketing spend, having the right influencers on board the project and targeting the correct marketing demographic.

- Before taking some influencers on board to promote the business, we can use datapoints such as retweets and the connected likes over a particular tweet for some other business to use other businesses as an example to understand an author's influence

- Given that most of the social media platforms provide us with demographic information, we can use our order history data to check if we had a spike in orders in the demographic where people liked a particular social media post. Depending on the order spike we can understand how well we are allocating our marketing spend

- For example: Loreal is coming up a new product line in Canada. They have already launched it in Europe and are looking to get an understanding of the Canadian market and the influencers they need to hire in order to have a successful launch

    - We can use twitter and tik-tok data points for other businesses to see how a product from another big market was accepted by the specific market population

    - We can also analyze how well certain influencers faired in the market by analyzing their reach and the opportunities they produced for other businesses.