TEXT ASSIGNMENT – TWITTER DATA SENTIMENT ANALYSIS

GOLD PRICES

Ashana Mehta

Roll no. 36

```
# Define the search term and the date_since date as variables
search_words = "gold+prices"
date_since = "2020-01-01"
```

```
# Collect tweets
tweets = tw.Cursor(api.search,
                   q=search_words,
                   lang="en",
                   since=date_since).items(1000)
tweets
```

After collecting keys and authenticating the same, the above code will fetch the results of **recent 1000 tweets** where the last date chosen is **1st January 2020** since there was a drastic change seen in gold prices suddenly from the beginning of this year due to recession and being considered as a “Safe Haven” asset, people seem quite positive about investing in the same.

```
#to remove retweets
new_search = search_words + " -filter:retweets"
new_search
```

```
'gold+prices -filter:retweets'
```

```
#get the new one (without re-tweets) in form of list
tweets = tw.Cursor(api.search,
                   q=new_search,
                   lang="en",
                   since=date_since).items(1000)

all_tweets = [tweet.text for tweet in tweets]
all_tweets[:7]
```

```
['Investors should be putting their money in gold now, as it represents a "very good hedge" ahead of risk events such… https://t.co/2XL63eCVb0',
 '@Damnzzelle Probably not 😂 \n\nWith gold prices the way they are, now is not a fun time to buy heavier gold pieces.',
 '@SkillUpYT I really fell for you guys, as much as we Americans complain about the prices of games we have it better… https://t.co/J86VTMcg44',
 'Gold prices climb back above $1,900 for highest finish in a week https://t.co/P7Fx1mHeMm',
 'Bitcoiners $BTC people \n\nWhy does bitcoin sell off when the market sells off? \n\nShouldn't it be nearly separate?… https://t.co/qj8o171ewZ',
```

Further, the re-tweets have been removed to conduct accurate analysis required for this topic and then the remaining actual tweets were been fetched in the form of a list.

```
#Details of users tweeting about gold prices
tweets = tw.Cursor(api.search,
                           q=new_search,
                           lang="en",
                           since=date_since).items(1000)

users_locs = [[tweet.user.screen_name, tweet.user.location, tweet.created_at,tweet.user.url] for tweet in tweets]
users_locs[:5]
```

```
#Create dataframe from the list of twitter data derived
tweet_text = pd.DataFrame(data=users_locs,
                          columns=['User', "Location", "Date & Hour of Tweet","URL Link"])
tweet_text[:5]
```

	User	Location	Date & Hour of Tweet	URL Link
0	NoreenAmore	Bangkok, Milan, Kuala Lumpur	2020-09-30 03:27:53	None
1	logical_blonde	Midwest	2020-09-30 03:27:36	None
2	TrueCross77		2020-09-30 03:22:23	None
3	playdoegold	London, Ontario	2020-09-30 03:21:31	None
4	FinanceBubble		2020-09-30 03:13:44	None

Here, important or required details were fetched from the tweets and transformed into a data-frame. Separate analysis could also be performed on the above data.

```
#to remove urls, hyperlinks, any mentions and replace with blank space
def remove_url(txt):
    return " ".join(re.sub("([^0-9A-Za-z \t])|(\w+:\/\/\S+)", "", txt).split())
```

```
#Call the above defined function in the list of tweets to create list of cleaned tweets
all_tweets_no_urls = [remove_url(tweet) for tweet in all_tweets]
all_tweets_no_urls[:5]
```

```
['Investors should be putting their money in gold now as it represents a very good hedge ahead of risk events such',
 'Damnzzelle Probably not With gold prices the way they are now is not a fun time to buy heavier gold pieces',
 'SkillUpYT I really fell for you guys as much as we Americans complain about the prices of games we have it better',
 'Gold prices climb back above 1900 for highest finish in a week',
 'Bitcoiners BTC people Why does bitcoin sell off when the market sells off Shouldnt it be nearly separate']
```

```
#Changing all into lower case - mentioned for loop as lower() doesnt apply on list
lower_case = [word.lower() for word in all_tweets_no_urls]
lower_case[:8]
```

```
['investors should be putting their money in gold now as it represents a very good hedge ahead of risk events such',
 'damnzzelle probably not with gold prices the way they are now is not a fun time to buy heavier gold pieces',
 'skillupyt i really fell for you guys as much as we americans complain about the prices of games we have it better',
 'gold prices climb back above 1900 for highest finish in a week',
```

Here, firstly after filtering out the re-tweets, all the hyperlinks and URLs were also been removed and all the tweets were been transformed into lower case in order to fetch good output in the “word cloud” to be formed and also to apply further analysis on the same.

```
#Check the word freqency by using split that splits words into unique elements of tweets
words_in_tweet = [tweet.lower().split() for tweet in all_tweets_no_urls]
words_in_tweet[:2]
```

```
# List of all words across tweets
all_words_no_urls = list(itertools.chain(*words_in_tweet))

# Create counter for each word as to how many times it has been used
counts_no_urls = collections.Counter(all_words_no_urls)

counts_no_urls.most_common(15)
```

```
[('gold', 958),
 ('prices', 676),
 ('the', 457),
 ('to', 304),
 ('and', 276),
 ('in', 236),
 ('of', 221),
 ('a', 203),
 ('silver', 183),
 ('for', 154),
 ('is', 153),
 ('as', 148),
 ('on', 134),
 ('you', 124),
 ('at', 117)]
```

Before applying word cloud and sentiment analysis, we need to clean the tweets in terms of stopwords and some unimportant words which would not add any value to our analysis. Thus, here firstly, the tweets are split into unique words and word-frequencies of top 15 words have been derived. This derived data could also be transformed into another data-frame for better visualization and further analysis.

For now, **top 25 words along with their word frequencies** have been plotted in the below horizontal bar charts.
But, since the top 25 words found did not seem to be useful, we further decide to **remove stopwords and collection words** from these words and then again plot a similar graph with new and important words along with their respective frequencies.

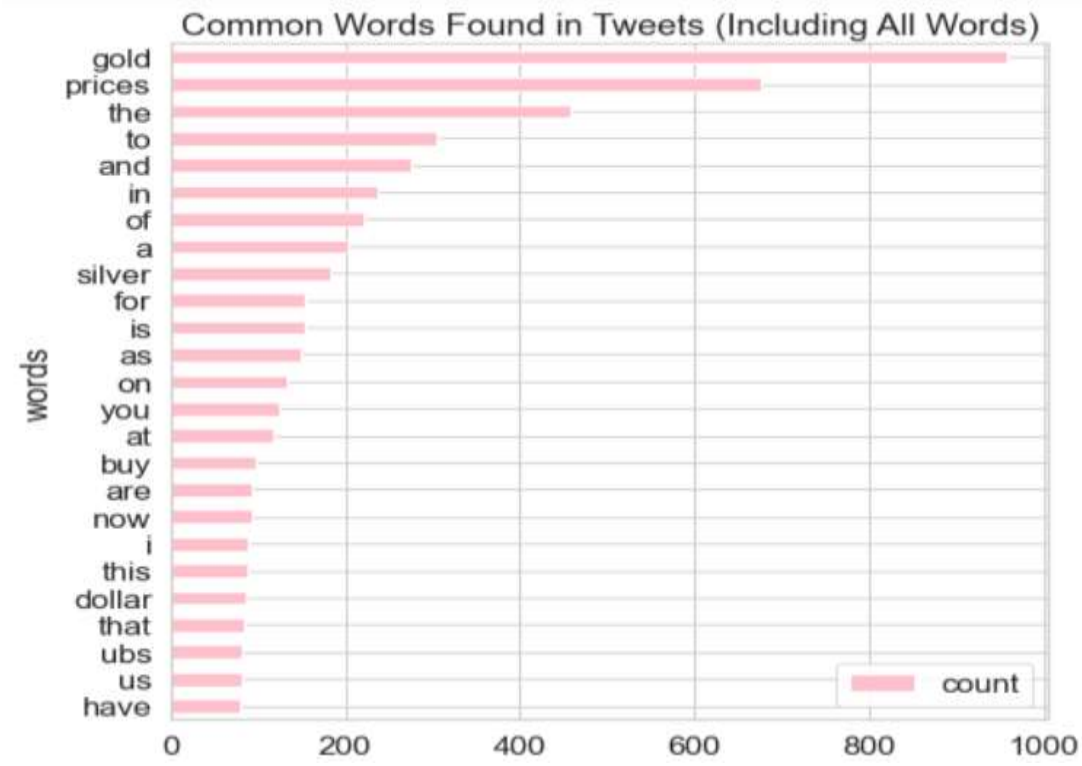
```
#Removed stopwords from top 25 words
stop_words = set(stopwords.words('english'))
```

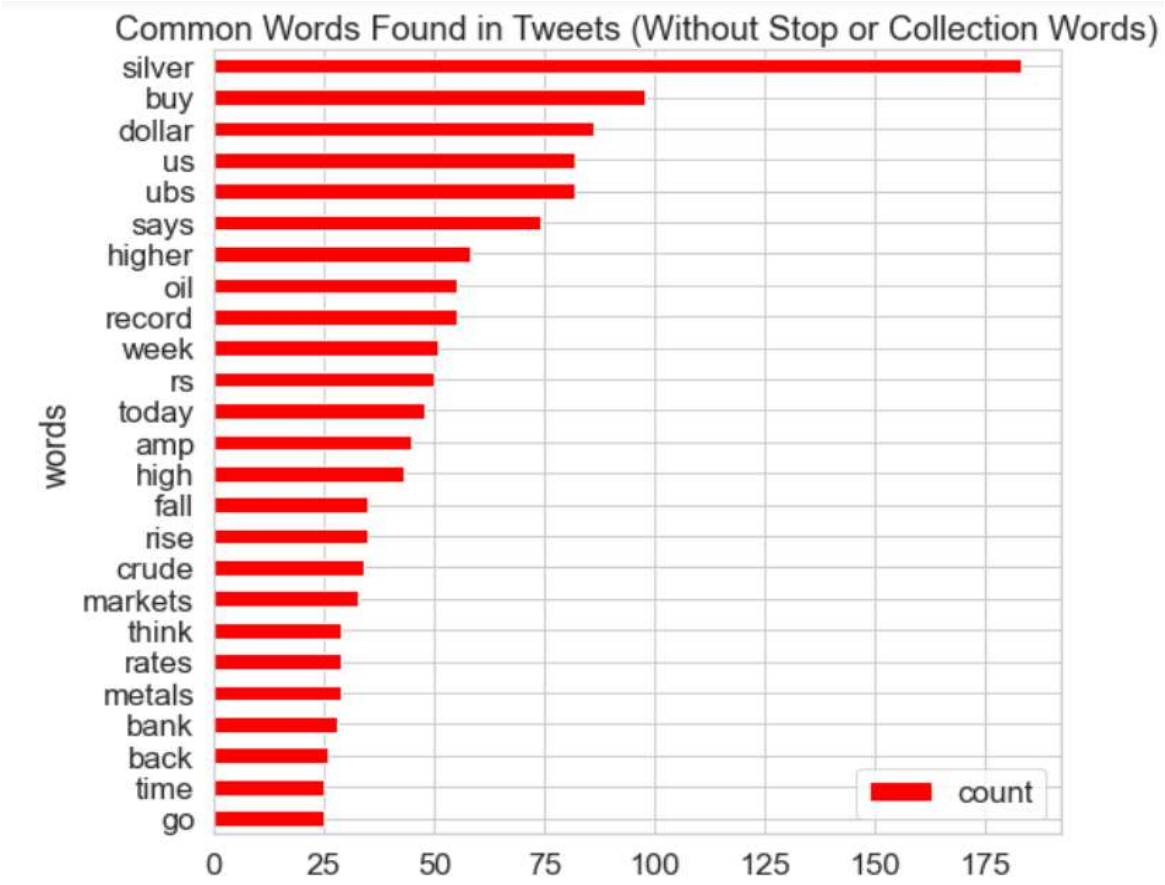
```
# Remove stop words from each tweet and get list of words
tweets_nsw = [[word for word in tweet_words if not word in stop_words]
              for tweet_words in words_in_tweet]

tweets_nsw[0]
```

```
['investors',
 'putting',
 'money',
 'gold',
 'represents',
 'good',
 'hedge',
 'ahead',
 'risk',
 'events']
```

```
#Defining collection words to be removed and adding it with that list without stopwords
collection_words = ['gold','prices','price']
tweets_nsw_nc = [[w for w in word if not w in collection_words]
                 for word in tweets_nsw]
```





Here, in the 2nd graph, some new and interesting words are seen which includes the markets, silver, foreign exchange currencies like “dollar and rupees (rs.)”, buy etc. ; somewhere indicating a positive attitude of people for gold.

Now, a wordcloud has been formed by transforming **all the filtered tweets into a data-frame** for better visualization.

```
#Creating word cloud - making a new dataframe where all the tweets are included
word_cloud = pd.DataFrame(lower_case,columns=['Tweet'])
word_cloud.head()
```

	Tweet
0	investors should be putting their money in gol...
1	damnzzelle probably not with gold prices the w...
2	skillupyt i really fell for you guys as much a...
3	gold prices climb back above 1900 for highest ...
4	bitcoiners btc people why does bitcoin sell of...

```
#Sentiment Analysis
from textblob import TextBlob
#Values closer to 1 indicate more positivity, while values closer to -1 indicate more negativity.
# Create textblob objects of the tweets
sentiment_objects = [TextBlob(tweet) for tweet in lower_case]
#getting polarity values of first tweet
sentiment_objects[0].polarity, sentiment_objects[0]
```

```
(0.45499999999999996,
 TextBlob("investors should be putting their money in gold now as it represents a very good hedge ahead of risk events such"))
```

```
#Transform it into a list from which a dataframe would be created
# Create list of polarity, subjectivity values and tweet text (3 things)
sentiment_values = [[tweet.sentiment.polarity, tweet.sentiment.subjectivity, str(tweet)] for tweet in sentiment_objects]

sentiment_values[0]
```

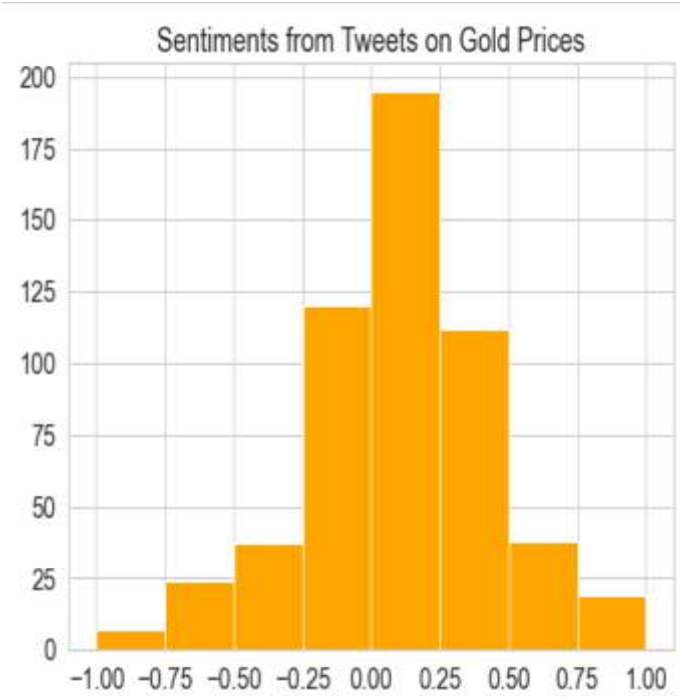
```
[0.45499999999999996,
 0.6400000000000001,
 'investors should be putting their money in gold now as it represents a very good hedge ahead of risk events such']
```

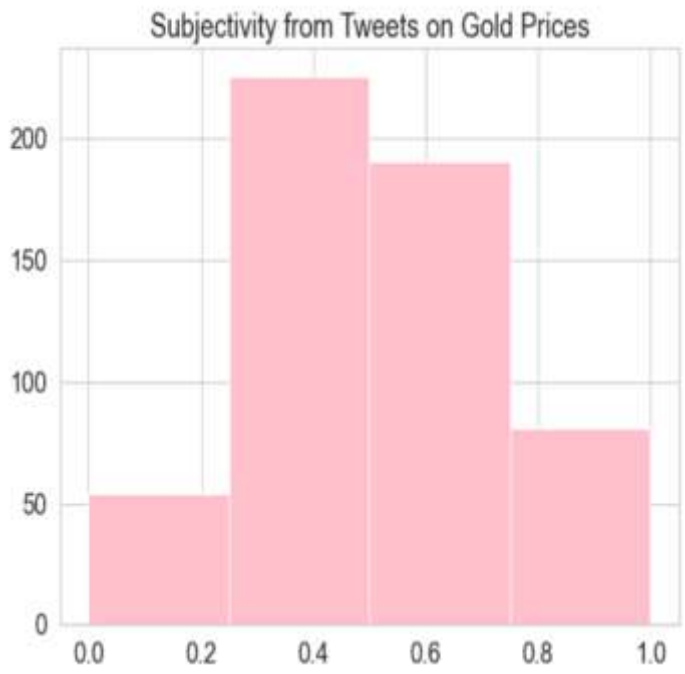
Lastly, **sentiment analysis** has been performed using **TextBlob** library and 'polarity' and 'subjectivity' values have been derived for the first tweet value mentioned above.
Further, a data frame has been formed with all the tweets with their respective polarity and subjectivity values derived. Also, here all the polarity values equal to 0 have been removed so as to get more accurate histogram of the same that could be easily visualized.

```
#Plotting histogram
# Remove polarity values equal to zero
sentiment_df1 = sentiment_df[sentiment_df.Polarity_values != 0]
sentiment_df1.head()
```

	Polarity_values	Subjectivity Values	Tweet
0	0.455	0.640000	investors should be putting their money in gol...
1	-0.150	0.200000	damnzzelle probably not with gold prices the w...
2	0.300	0.300000	skillupyt i really fell for you guys as much a...
4	0.100	0.400000	bitcoiners btc people why does bitcoin sell of...
5	-0.150	0.133333	gold last week recorded its biggest weekly dro...

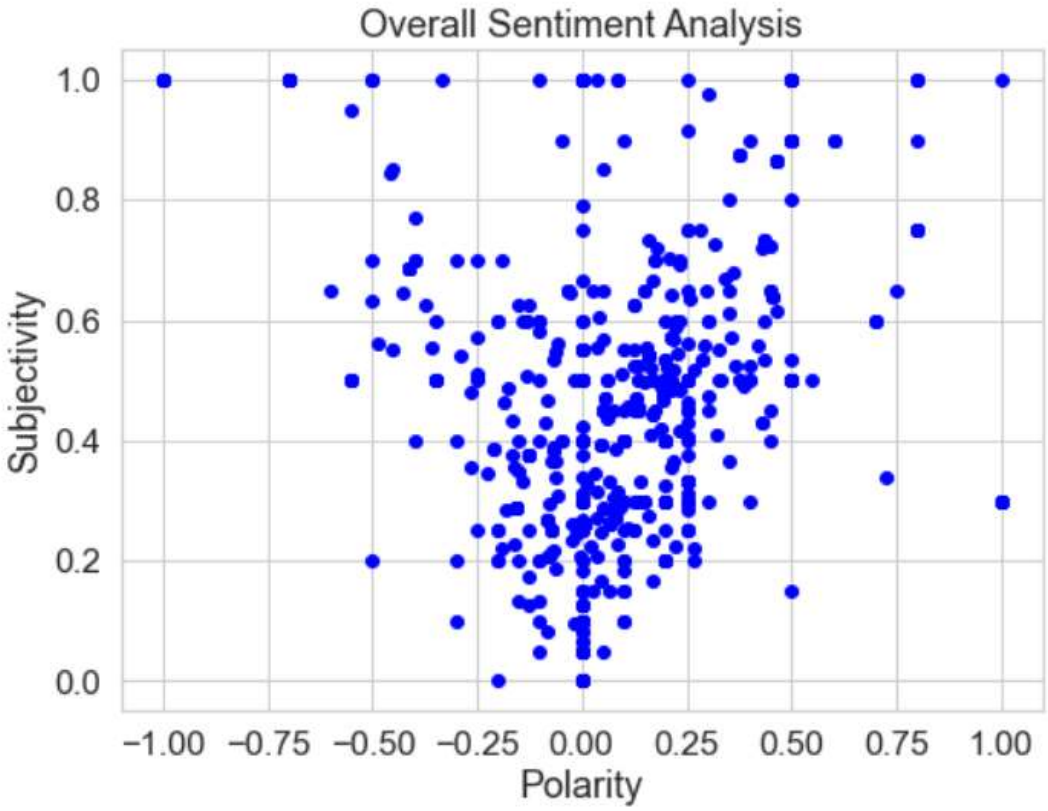
Below are the histogram plots for the polarity and subjectivity values respectively.





Here, it is clearly seen that the polarity values are more inclined towards 1, thus indicating positive sentiment of people for gold prices as there had been a great surge in the prices of the same during this pandemic. At the same time, subjectivity values are also falling more in the range of **0.4 to 0.8**, indicating that there exists considerate amount of subjectivity in the tweets i.e. most of the tweets are highly opinionated.

For better understanding of the same, polarity values (with 0s) and subjectivity values have been plotted in a “Scatter Plot” in order compare the same.



Here, it is clearly seen that most of the polarity values that fall towards the extremity i.e. 1 and -1; have its subjectivity as 1, indicating that any kind of **strong positive or negative sentiments shown in the tweets mentioned by the user were highly opinionated to certain extent.** Rest of the other tweets represented having positive sentiments with moderate to low subjectivity.