Project Deliverable 3

Group Members: Conor Walsh, Patrick Kuzdzal, Yash Bengali, Carlos Lopez

---

Refine the preliminary analysis of the data

We added a few more data points in our preliminary analysis of the data. We made a column which contains all the open, closing, low, and high points of the data from yahoo finance. We also added a column that has the closing-opening price and a column that is 0 or 1 based on if the stock price overall increased or decreased on that day. We also used Google Cloud's Natural Language Processing API to gather sentiment scores for around eighty thousand tweets. In addition we used the NLTK package to conduct sentiment analysis on the entire kaggle dataset.

Answer another key question

Is there a correlation between stock volume and number of tweets per day?

There does appear to be at least a moderately strong correlation between the number of tweets containing a specific stock name, and the trading volume of that stock on that day. This was found out through counting the amount of tweets that were able to be scraped on a specific trading day, and graphing both that and the trading volume data for that day that was acquired through Yahoo Finance. Indeed, after normalizing both pieces of data by dividing each datapoint by the max element in the corresponding dataset, a correlation of 62.8% was found in the tweets within the kaggle dataset, and an even higher 84% correlation was found between the tweets that we scraped and the stock's corresponding daily trading volume.

Attempt to answer overarching project question

In terms of the relations between the sentiment of tweets and stock price, there were some promising correlations, detailed more explicitly in the 'Results' section.

On the side, we found a correlation between the stock volume and number of tweets, which shows that there is at least some connection between tweets and stocks, even if it is just a reactionary relation.

Refine project scope, list of limitations, and potential risks of achieving project goal

Our aim is still to find a correlation between Apple's stock price and the general sentiment for that company on Twitter per day. That said, as is explained above, there were some discrepancies with achieving this goal. We are now planning on expanding to other companies beyond $AAPL to see if a correlation holds, namely $TSLA, $AMZN, and or $GOOG.

In terms of limitations with data, one impactful bound is the timeframe that the Twitter API limits users to, as developer accounts are only able to scrape tweets from the past week. We

attempted numerous proposed workarounds that sought to bypass this bottleneck by parsing the tweets as HTML, however to no avail, due to updates that Twitter rolled in early 2021. In the end, we adopted the limitation, and collected tweets totalling just under 80,000. The other main roadblock occurs with Google Cloud's Natural Language Processing API. This service carries a minor financial burden, in addition to a lengthy time-to-run, which makes it infeasible to run on the Kaggle dataset. That said, it has also produced the most promising results, so we chose to value it more than other means. Ultimately, we ran it on the manually collected dataset.

We do not see any potential risks of achieving our project goal, as we are just looking to determine whether a correlation exists between the sentiment of tweets about Apple, and their stock price.

Final Report Draft

Group Members: Conor Walsh, Patrick Kuzdzal, Yash Bengali, Carlos Lopez

---

Goal

The goal of this project was to determine if there exists a correlation between the sentiment of tweets regarding $AAPL on twitter, and the stock price on a given day. We examined the emotionality of tweets (i.e. how the amount of positivity/negativity of a tweet impacts stock price), pure sentiment scores, and a binary indicator of positive/negative day, in relation to change in stock price from open to close. Ultimately, the results found are detailed below.

Approach

The twitter api is limited to scraping tweets from the past 7 days. So while we scraped about a month of tweets on our own, we also found a kaggle dataset that contains tweets about big tech companies from the past 5 years. Our approach was as follows:

1. Get the tweets in a given time frame
2. Get the stock data from yahoo finance in a given time frame
3. Perform the sentiment analysis of the tweets
4. Run correlation analysis to check if the mean sentiment on a given day correlated to a related stock increase/decrease

The multiple different ways that tweets could be related to the stock price. One possible way is that twitter reflects the action of a stock in that if the stock goes up, then the tweets have a more positive sentiment than on days that the stock goes down. Another possible relation is that tweets are an indicator of potential movement, as in if tweets are really positive one day, the stock is likely to go up the next, and if the tweets are negative then the stock is likely to go down. Another thing to consider is if tweets are about Apple the company or Apple the stock. An example positive tweet about Apple the company is “I love apple, my iPhone is amazing”, although overwhelmingly positive this tweet is not likely to relate to a one day rise in price of Apple’s stock, rather a long term positive trend. Although analyzing such a long term trend would be difficult due to surrounding factors such as the economy that we would have to account for. Related to that, scrapping for purely Apple tweets is somewhat difficult due to apple being a word as well, in addition many tweets about Apple not even including the word Apple, for example someone could tweet about their iPhone without mentioning Apple. Although, this might be useful if someone was purely interested in finding a long term public feel for Apple is by looking only at Apple’s products because you would not get any tweets about apple the food or people tweeting about Apple’s stock. Due to these factors, we decided to focus on determining if there is a correlation between tweets and an individual day’s stock movement. In addition, due

to the limits of the twitter api, the majority of tweets are from a kaggle dataset that uses Apple's ticker $AAPL to identify Apple tweets.

Results

When analyzing a smaller section of tweets that have a sentiment score from Google Cloud's api, we found that there was a significant positive correlation, a Pearson r value of .613 and a p value of .019. Although when using a different sentiment analysis on a larger dataset when found that there was a slight negative correlation, a Pearson r value of -0.51. We also found that there was no correlation between the absolute value of sentiment of the tweets and the price of the stock. We tested this because it was possible that on days that the stock went up there were stronger reactions, both positive and negative. Although we did find that there was a negative correlation between the sentiment of tweets and the magnitude of the change, r is -0.68, p is .0067. This indicates tweets tend to be more positive on days that the stock does not change much.

In addition to only analyzing the price of the stock, we analyzed the volume of stock. Specifically we compared the stock volume and number of tweets per day. We found that there was a positive correlation, r is .62.

Visualization

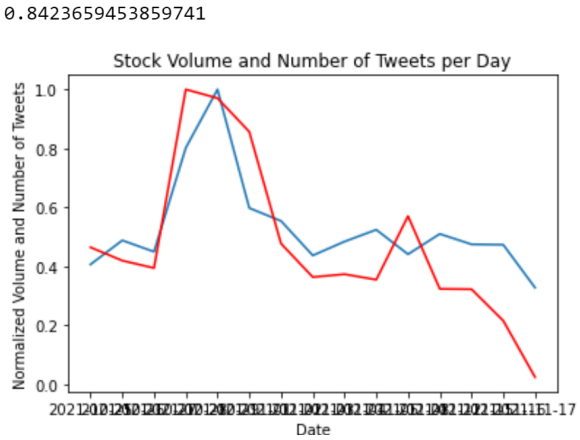

Figure 1: 84% Correlation between Stock Volume and number of Tweets per Day

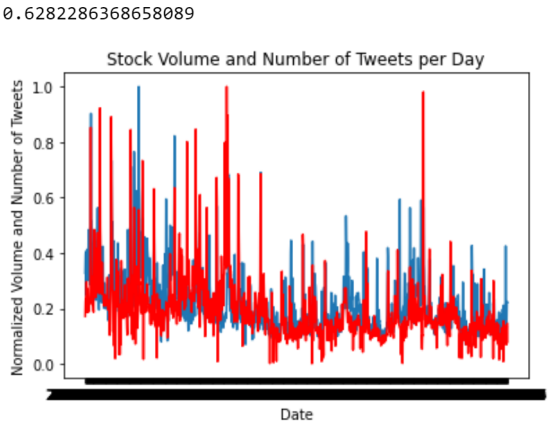

Figure 2: 63% Correlation between Stock Volume and number of Tweets per Day

```
          date    change     SCORE
1   2021-10-25 -0.039994  0.048479
2   2021-10-26 -0.009995  0.050156
3   2021-10-27 -0.509995  0.043779
4   2021-10-28  2.750000  0.050630
5   2021-10-29  2.580002  0.048773
8   2021-11-01 -0.029998  0.040465
9   2021-11-02  1.360000  0.047538
10  2021-11-03  1.100006  0.043370
11  2021-11-04 -0.619995  0.040230
12  2021-11-05 -0.610000  0.036411
15  2021-11-08 -0.970002  0.044341
16  2021-11-12  1.560012  0.049297
19  2021-11-15 -0.369995  0.042587
20  2021-11-16  1.059998  0.042389
           SCORE    change
SCORE   1.000000  0.613123
change  0.613123  1.000000
(0.613122970995022, 0.019722687543916626)
```

Figure 3: Correlation between the mean sentiment score for tweets and the change between close and open for that day.

```
          date    change     SCORE
1   2021-10-25  0.039994  0.050366
2   2021-10-26  0.009995  0.056344
3   2021-10-27  0.509995  0.040897
4   2021-10-28  2.750000  0.023406
5   2021-10-29  2.580002  0.035297
8   2021-11-01  0.029998  0.046025
9   2021-11-02  1.360000  0.052338
10  2021-11-03  1.100006  0.045253
11  2021-11-04  0.619995  0.053475
12  2021-11-05  0.610000  0.044241
15  2021-11-08  0.970002  0.040185
16  2021-11-12  1.560012  0.051759
19  2021-11-15  0.369995  0.047174
20  2021-11-16  1.059998  0.047948
           SCORE    change
SCORE   1.000000 -0.685893
change -0.685893  1.000000
(-0.6858929544235952, 0.00676469676106771)
```

Figure 4: Correlation between the mean sentiment score for tweets and the absolute value for the difference between close and open for that day.

```
          date    change     SCORE
1   2021-10-25 -0.039994  0.166536
2   2021-10-26 -0.009995  0.162630
3   2021-10-27 -0.509995  0.165845
4   2021-10-28  2.750000  0.160385
5   2021-10-29  2.580002  0.163316
8   2021-11-01 -0.029998  0.158751
9   2021-11-02  1.360000  0.158617
10  2021-11-03  1.100006  0.156995
11  2021-11-04 -0.619995  0.168119
12  2021-11-05 -0.610000  0.156254
15  2021-11-08 -0.970002  0.161395
16  2021-11-12  1.560012  0.163877
19  2021-11-15 -0.369995  0.157343
20  2021-11-16  1.059998  0.170190
           SCORE    change
SCORE   1.000000 -0.010681
change -0.010681  1.000000
(-0.010681079700207596, 0.9710914804968317)
```

Figure 5: Correlation between the mean of absolute values of the sentiment scores for tweets and the change between close and open for that day.

Code

As for the code, we first run the python file titled ‘tweet_scraper’, which utilizes the official Twitter API to scrape tweets with the keyword $AAPL from the last 7 days. This script also converts all of the tweets into a set, so as to remove duplicates, and writes them to a file named ‘tweets.csv’. In said file, each row follows the order: date of creation, number of followers the parent account has, number of likes the post received, and the content of the tweet itself.

The second main data collection script is named ‘gcloud_sentiment’, which reads the tweets from the aforementioned file, and sends them to Google Cloud’s Natural Language Processing API for sentiment analysis. Returned for each post is a score and magnitude, which represents its emotional nature on a scale of -1 to 1. These values are written into a file named ‘sentiments_new.csv’, which follows the order: score, magnitude.

One thing to note is that both of these scripts require their respective credentials to function, existing in the form of a ‘credentials.json’ and ‘creds.json’ file, however these were not included for obvious security based reasons.

To find the correlation between we put the tweets, apple stock prices and sentiment scores into panda dataframes. We then combine tweets and sentiment scores so that every tweet has a sentiment score. After dropping unnecessary columns, we group the combined tweets dataframe by data, calculating the mean score for that data. To get the difference between the close and open, we created a new column in the dataframe and removed the unnecessary columns. Then we join that dataframe and the apple stock prices dataframe on the data and then use panda’s .corr() function to get the Pearson r value and scipy.stats to get the Pearson r value again with the corresponding p value. If we are finding the correlation to something slightly different, we just modify the column using the apply function and a lambda function, for example, making all the scores positive. This code is in a python notebook file called ‘correlation.ipynb’ in the ‘google-test’ branch of the repo.

Source

[1] https://www.kaggle.com/omermetinn/tweets-about-the-top-companies-from-2015-to-2020?select=Company_Tweet.csv