Final Deliverable

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.

In terms of actually getting the sentiment of the tweet there are a couple of different approaches that we tried. The first is to use Google Cloud's Natural Language Sentiment Analysis api. This would probably give us the most accurate results in terms of actually classifying tweets as generally positive or negative although it has limitations. Another approach to sentiment analysis that we took was to use a dataset that contained stock related tweets that were labeled as positive or negative[2]. Then train a model to classify the tweets as positive or negative, and feed in the new tweets to get a sentiment score. This approach of sentiment analysis is much more targeted towards our use case.

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.

When using the stock specific sentiment analysis, we found that there was no significant correlation. We tried many different combinations, shifting the days to see if the tweets were reacting a day late, or as an indicator. We also tried only looking at more recent years, both these resulted in no improvement to the correlation. The only thing that resulted in even a marginal increase was to limit the tweets to notable tweets, i.e. tweets that had a certain amount of likes or comments, although it still is not significant. Something to note is that out of the hundreds of thousands of tweets, there are only 1000 tweets with over 50 likes and 5 comments, and only 200 with over 500 likes and 5 comments, meaning that the tweets that we are looking at are not that influential or widely viewed.

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

```
0.8423659453859741
```

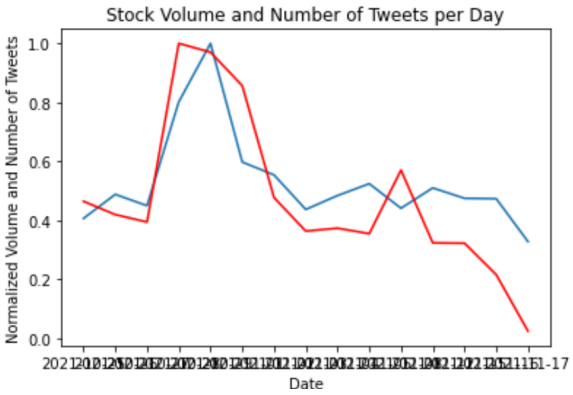

Figure 1: 84% Correlation between Stock Volume and number of Tweets per Day

```
0.6282286368658089
```

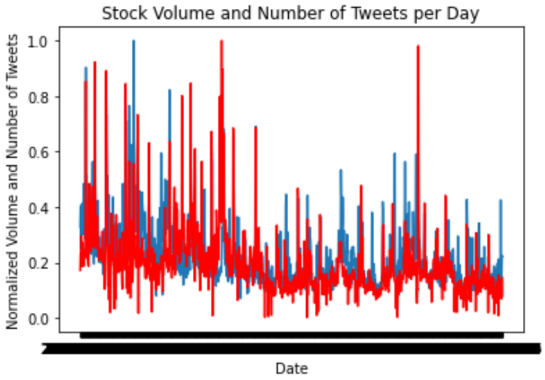

Figure 2: 63% Correlation between Stock Volume and number of Tweets per Day

	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.

	increased	change_value	Mean_Sentiment
increased	1.000000	0.711223	-0.039479
change_value	0.711223	1.000000	-0.063487
Mean_Sentiment	-0.039479	-0.063487	1.000000

Figure 6: Correlation if the stock increased that day and the change between the close and open for that day with the mean sentiment of tweets for that day. Using the kaggle tweets dataset which spans between 2014 and 2020.

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.

The code that used the alternative kaggle dataset to classify the tweets as positive or negative is in tweets_sentiment.ipynb. The labelled tweets are read into a dataframe and then processed. The processing takes in the dataframe and uses nltk to stem the words, and sklearn's TfidfVectorizer to vectorize the tweets. Finally it sklearn TruncatedSVD to reduce the number of features to a manageable number, we specifically used 60 because that was the most that we used given the limitations of our available computing power. Then we split the training data into training and testing sets and used xgboost's XGBRegressor to create a model that classifies the tweets. The model has a RMSE on the testing set of 0.2572 which is pretty good. Then all of Apple's tweets go through the same process that the training tweets went through, they are stemmed, vectorized, decomposed and given to the model that gives the tweet a sentiment score.

Challenges

Getting Tweets

One of our biggest challenges in this project was getting the tweets needed to analyze the data. We got twitter developer accounts and utilized tweepy to help scrape tweets. However, tweepy has a restriction that only tweets from the past 7 days can be scraped. Therefore, we started looking elsewhere to find data and eventually found a kaggle dataset which contained tweets about large companies over the last 5 years. We proceeded to use that dataset until we were able to scrape enough tweets on our own to migrate to our own dataset. This data was of the past month whereas the kaggle dataset was from 2015 to 2020.

Scraping tweets using $ vs

Tweets normally are categorized under a hashtag. Although this makes it easier to group tweets, we noticed that many of the tweets we scraped with # were just retweets, and many were not relevant. We also looked at using $ which contains tweets related to a specific company. For example, $AAPL. This is used to categorize tweets under apple which are more relevant towards the stock data and towards the company. Usually the tweets are related to more towards the stock data and we decided to proceed down this route as we thought it would prove to have higher correlations.

Differences in Sentiment Analysis

Our first approach consisted of analysing the tweets we had gathered up to that point with nltk's sentiment analysis tools. We thought this was a nice choice since it offered tokenizing, topic segmentation, stemming, and a variety of other tools that we could use for a more thorough sentiment analysis. In the library specifically, we tried to use the NLTK sentiment intensity analyzer tool. The NLTK sentiment intensity analysis library which returns back a negative, neutral, positive, and compound score. We chose to use the compound score as it was an aggregate of the other scores. We grouped the tweets by the days and computed the compound analysis for each of these tweets and averaged them over the day. This resulted in a 11% correlation when comparing the compound sentiment scores over a day and whether the stock increased or decreased.

This issue was resolved when we dabbled in a few other sentiment analysis libraries and landed upon the one provided by the Google Cloud API. After re-running the same tests we indeed arrived at a much better model for predicting stock prices since it had found a much higher correlation between the average Twitter sentiment and the stock price of Apple for that day. Hence, this leads us to believe that the sentiment analysis that is provided by the Google Cloud API is more thorough than other competitors like nltk and Spacy (which makes sense considering that the Google Cloud API is not free).

When using the kaggle dataset that contains labeled tweets, there are a couple of limitations. One of them being that the dataset only had 1300 labeled tweets, and they were manually labelled without clear guidelines to what constitutes a positive or negative tweet, especially since there seems to be many just neutral tweets. That aside, we used tfidf vectorization to convert the text to something that the model can use, and although we only used 60 features, it took over 14 hours to convert the text. Perhaps using more columns could have led to more accurate results.

Code for non-chosen approaches:

test.ipynb contains our code for analyzing the code using nltk intensity analyzer. Our results can be recreated by running the jupyter notebook if desired.

Conclusion

It appears that despite it just being a social media platform, data gathered from Twitter can effectively gauge the general public's interest in a stock for a particular trading day. For example, the amount of Tweets posted in a day, heavily correlated with the trading volume for that day. Indicating a direction for the stock for that particular day seems to be a more convoluted process. Additionally, we noticed that the number of tweets has a negative correlation with stock price. Hence, although this should not be used as an "end-all" decision for when to trade Apple stock, it can definitely be used as an indicator for when a possible trading opportunity may arise.

Source

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