MATH189 Final Project

Jeremy Tsai

May 2020

1 Research Topic/Motivation

My main research question centers around if social media pertaining to COVID-19 can be used to accurately gauge the perception/emotion of the general public. If social media platforms carry mostly negative sentiments for a given period of time, does this cause people to react in a certain way? How about versus when social media platforms carry mostly positive sentiments? Specifically, I have chosen to analyze the sentiments of Tweets related to COVID-19 and their correlation to the closing prices of three major stock market indices: the Dow Jones Industrial Average, the NASDAQ Composite, and the Standard and Poor's 500 (S&P500). To conduct sentiment analysis on the Tweets, I utilized the VADER Sentiment Analysis package from the Natural Language Toolkits (NLTK) library. I was motivated to pursue this research topic because Natural Language Processing has always been an interesting application of Machine Learning. Furthermore, I wanted to understand the power of social media on the public and whether it could have an impact even on something as big as the stock market.

2 Past Research

Several studies have been done in the past about the power of Tweets to influence stock markets. One especially interesting study was "The Impact of Donald Trump's Tweets on Financial Markets" by a student from the University of Nottingham. Given Trump's role as the President of the United States and his massive following of over 50 million individuals on Twitter, it makes sense that his Tweets would have an influence on people and even the stock market. This study also utilized VADER sentiment analysis (which is talked about more in depth in the next section) in order to assess how positive/negative each of Trump's Tweets were. In the end, they found that Trump can indeed influence financial markets with his Tweets. Both positive and negative Tweets led to significant abnormal returns, with the effect depending on whether the Tweet was positive or negative (positive Tweets led to positive returns and negative Tweets led to negative returns). They found that positive Tweets also had more effect than negative Tweets.

3 VADER Sentiment Analysis

To classify the daily Tweets I gathered, I decided to utilize VADER Sentiment Analysis. Sentiment Analysis in general refers to using Natural Language Processing to identify and extract opinions and emotions from a given text. In this age of Big Data, Sentiment Analysis allows companies and individuals alike to efficiently sift through massive amounts of information and gain key insights that can benefit them. In my search for sentiment analyzers to use, I came across VADER, which fit my needs perfectly. VADER (Valence Aware Dictionary and Sentiment Reasoner) is a polarity-based and valence-based sentiment analyzer. This means it classifies texts as either positive or negative but also outputs the intensity of the sentiment. Using the analyzer and passing in a piece of text, it will return a compound score between -1 and 1, with -1 signifying the text is extremely negative and 1 signifying the text is extremely positive. It works especially well on social media text since it can handle emojis, emoticons, and slang. Furthermore, it pays close attention to capitalization which can denote varying levels of emotion ("great" vs. "GREAT"), conjunctions (like "but") that can shift the sentiment of a text, and degree modifiers (like "extremely" vs. "marginally").

The way that VADER Sentiment Analysis works is having a lexicon of words that are related to sentiment. Within this lexicon, each word is given a rating to denote how positive or negative it is. To make this accurate, the creators of VADER actually utilized Amazon Mechanical Turk, a micro-labor website where people can be hired to perform small tasks. Obviously, the chosen raters had to pass specific tests to ensure that their ratings would be of high quality. When given some text, the VADER analyzer looks through it and checks to see if any words exist in the lexicon. The ratings of the words present are then used to calculate the compound sentiment score between -1 and 1. Of course, this is not the only thing that VADER does. To be even more successful in sentiment analyzing, it analyzes the word context by paying attention to capitalization, grammar, modifiers, and syntax.

4 Methods

4.1 Data Collection

To begin answering the main question of whether the sentiments of Tweets regarding COVID-19 had an impact on the closing prices of the three major stock market indices, I had to gather Tweets related to COVID-19 and closing prices of the indices (with corresponding dates). I obtained the daily closing prices of the Dow Jones Industrial Average, S&P500, and NASDAQ Composite between 3/12/2020 and 4/30/2020 from Yahoo Finance. Additionally, I gathered daily CSV files containing millions of Tweets between 3/12/2020 and 4/30/2020 from Kaggle.

4.2 Data Processing

The data from Yahoo Finance regarding the indices came with a lot of extra information such as opening price, high price, and low price. I deleted all of these columns and kept the date and closing price columns. The daily COVID-19 Tweet data from Kaggle also contained

lots of extraneous columns like screen name and retweets. I deleted these columns and kept the date, text, and language columns. I had to keep the language column because the next step of data processing was filtering out Tweets that were not in English. The language column had "en" for Tweets that were English, making the filtering straightforward.

4.3 Sentiment Analysis

To conduct the sentiment analysis, I created an instance of the VADER sentiment analyzer and had it loop through the text of all Tweets for a given day. The compound sentiment score of all Tweets for a day would be added up and then divided by the total number of Tweets for that day to gain the average sentiment rating. The average sentiment rating of each day was then recorded.

4.4 Correlation Analysis

After calculating the sentiment rating of each day between 3/12/2020 and 4/30/2020, I plotted them against the daily index closing prices on 3 separate graphs, 1 for each of the major indices. Then, using the scipy library, I conducted a linear regression to see how correlated the two variables were and the significance of the correlation.

5 Results

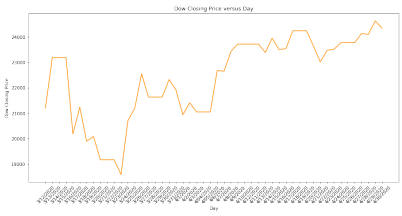

Figure 1: Dow Closing Prices vs. Day

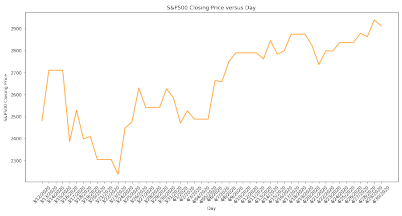

Figure 2: S&P500 Closing Prices vs. Day

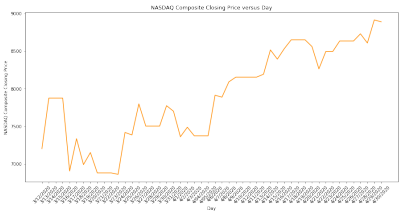

Figure 3: NASDAQ Closing Prices vs. Day

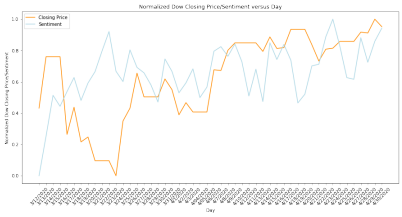

Figure 4: Normalized Dow Closing Prices/Sentiment vs. Day

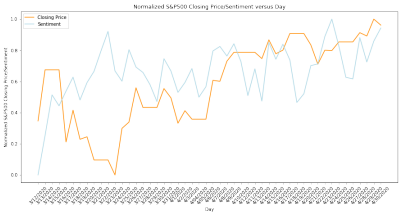

Figure 5: Normalized S&P500 Closing Prices/Sentiment vs. Day

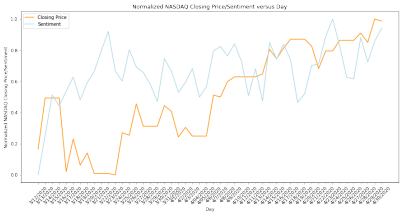

Figure 6: Normalized NASDAQ Closing Prices/Sentiment vs. Day

```
p-values: 0.13972463308999897
R^2: 0.04487319773132123
Slope: 19366.530389048025
```

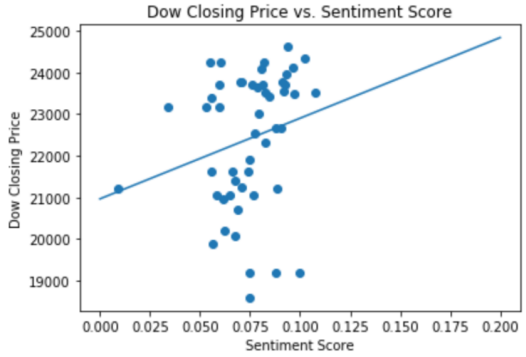

Figure 7: Dow Closing Prices vs. Sentiment

```
p-values: 0.06868556114182187
R^2: 0.0673857206519471
Slope: 2767.8720301441303
```

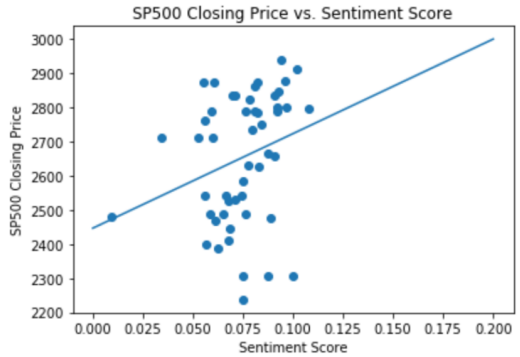

Figure 8: S&P500 Closing Prices vs. Sentiment

```
p-values: 0.013248091009335442
R^2: 0.12115817152002392
Slope: 12105.706217976827
```

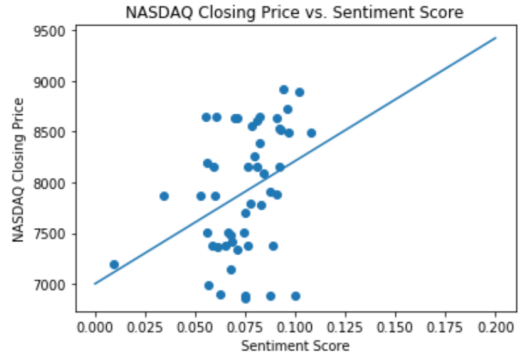

Figure 9: NASDAQ Closing Prices vs. Sentiment

In Figures 1-3, we see the plots of each of the major indices plotted against days. In Figures 4-6, we see the plots of the normalized index closing prices in orange and the normalized sentiment scores in blue plotted against days. This is to help see how the two variables tracked with each other over time.

To mathematically quantify the correlation between closing prices and sentiment scores, we look at Figures 7-9. In Figure 7, we see daily Dow closing prices plotted against their corresponding daily average Tweet sentiment scores. The correlation coefficient is 0.045 and the p-value is 0.140. Using a significance threshold of $p = 0.05$, we do not have a significant relationship between the two variables. In Figure 8, we see daily S&P500 closing prices plotted against their corresponding daily average Tweet sentiment scores. The correlation coefficient is 0.067 and the p-value is 0.069. Using a significance threshold of $p = 0.05$, we do not have a significant relationship between the two variables. In Figure 9, we see daily NASDAQ closing prices plotted against their corresponding daily average Tweet sentiment scores. The correlation coefficient is 0.121 and the p-value is 0.013. Using a significance threshold of $p = 0.05$, we find a significant relationship between the two variables. Below is a table summarizing the results.

	Correlation Coefficient	**p-value**	**Significiant?**
Dow	0.045	0.140	No
S&P500	0.067	0.069	No
NASDAQ	0.121	0.013	Yes

Table 1: Significance findings for correlation between sentiment scores and closing prices

6 Discussion

From Table 1, we see that there was only a significant positive correlation between average daily Tweet sentiment scores and the NASDAQ Composite closing prices. For the Dow and S&P500 indices, although their closing prices visually seemed positively correlated with the sentiment scores, their p-values were slightly too high to declare significance.

These results only partly fit my initial prediction. I hypothesized that there would be a positive correlation between sentiment scores and closing prices. However, only the NASDAQ closing prices proved to have a significant relationship with the Tweet sentiment scores. I would think that this is because if sentiment is higher, the general public feels more secure and stable, thus reflecting over to higher prices in the stock market. However, if this were the case, I would expect significance for all three indices instead of just the NASDAQ.

Furthermore, we must remember that correlation does not imply causation. There is the directionality problem as well as the third variable problem. We do not know if the closing prices are impacting the sentiment scores or vice versa. Maybe positive Tweets cause increased investment confidence or higher stock prices cause people to Tweet more positively. We cannot conclude which direction is correct based on the steps of this study. There could

also be outside variables that we did not account for which are influencing both closing prices and sentiment scores. Additionally, in this study, we only looked at the sentiments of Tweets related to COVID-19. Although COVID-19 is probably the most-talked about topic on Twitter right now, there are various other subjects that people are Tweeting about. If we included these Tweets in our sentiment analyses, there is no doubt that the average daily Tweet sentiment scores would be altered along with the correlations we found.

There are various extensions of this research that can be done. This study only looked at the closing prices between 3/12/2020 and 4/30/2020. Opening prices and volume could also be examined, along with an extended time period. Sentiment analyzers different from VADER could be tested to see if they yield the same correlation results. We could also expand the Tweets that we analyzed to include those that are not related to COVID-19. We could even try to go beyond looking at Tweets from Twitter but other social media platforms like posts from Instagram and Facebook. There is clearly still much to be learned about the connection between social media sentiment and the financial markets!

7 References

1. Past Research on Tweets and the Stock Market:

   `https://www.nottingham.ac.uk/economics/documents/research-first/krishan-rayarel.pdf`

2. Dow Jones Closing Prices:

   `https://finance.yahoo.com/quote/%5EDJI/history?period1=1583971200&period2=1588291200&interval=1d&filter=history&frequency=1d`

3. NASDAQ Composite Closing Prices:

   `https://finance.yahoo.com/quote/%5EIXIC/history?period1=1583971200&period2=1588291200&interval=1d&filter=history&frequency=1d`

4. S&P500 Closing Prices:

   `https://finance.yahoo.com/quote/%5EGSPC/history?period1=1583971200&period2=1588291200&interval=1d&filter=history&frequency=1d`

5. COVID-19 March Tweets:

   `https://www.kaggle.com/smid80/coronavirus-covid19-tweets`

6. COVID-19 Early April Tweets:

   `https://www.kaggle.com/smid80/coronavirus-covid19-tweets-early-april/data`

7. COVID-19 Late April Tweets:

   `https://www.kaggle.com/smid80/coronavirus-covid19-tweets-late-april`

8. Vader Sentiment Analysis

   http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf

9. Vader Sentiment Analysis:

   http://t-redactyl.io/blog/2017/04/using-vader-to-handle-sentiment-analysis-with-social-media-text.html

10. Vader Sentiment Analysis:

    https://medium.com/analytics-vidhya/simplifying-social-media-sentiment-analysis-using-vader-in-python-f9e6ec6fc52f