Final Project Progress

Jeremy Tsai

**Changes:**

I have had to slightly change the focus of my project due to difficulty in acquiring recent data regarding food expenditures, oil consumption, and overall consumption. Thus, I have shifted the focus to the changes in the most popular stock market indices: Dow Jones, S&P500, and NASDAQ Composite. I will continue to do sentiment analysis of tweets related to COVID-19 by day and see if there is any correlation between the positivity/negativity of the tweets in a day and the rise/fall of the closing prices of the three aforementioned indices. If there is a correlation, then it could signify that the tweets can be used to accurately judge the public’s perception and that these tweets could have enough social influence to impact the stock market.

**Progress:**

For the data regarding the stock market indices, I have found the daily closing prices of the Dow Jones, S&P500, and NASDAQ Composite from the beginning of March to the end of April. The links are below. The CSV files are also located within the final project folder.

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

NASDAQ Composite:

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

S&P500:

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

For the data regarding tweets, I have found daily tweets from the beginning of March to the end of April. The links are below. The CSG files are also located within the final project folder. I will need to filter out the irrelevant information (only want to keep English tweets and text of the tweets).

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

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

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

I have already found a semantic analyzer called VADER and it is imported from the nltk library (Natural Language Toolkit). It is specialized for analyzing texts from social media. I will explain it more in depth in my final paper. For each day, I want to run the sentiment analysis model through all the tweets. The VADER sentiment analyzer outputs a compound score from -1 to 1, where 1 means the text is completely positive and -1 means the text is completely negative. I will sum up the sentiment scores of all the tweets within each day and store this data.

To combine the data regarding the closing prices of the three indices and the positivity/negativity of tweets for the days between March and April, I want to first plot the normalized stock market index closing price versus day for all three indices. Then, for each respective index, I will plot the data of positivity/negativity of each day over the normalized data points for the closing prices for each day. From this, we will be able to see if the positivity/negativity of the tweets for a day track the closing price of the index at that day. I would expect that when the tweets for a day are more negative, then the closing price will be lower versus another day that had more positive tweets. Furthermore, I will plot the closing price versus the positivity/negativity statistic for each day and conduct statistical tests to see if there is any correlation.

**Final Report Structure:**

1. Motivation
2. Past Research
3. Vader Sentiment Analysis
4. Methods
5. Results
6. Discussion