CMPT353-StockAnalysis Report

Timothy Tan
Ernest Wong
Tai Luan Nguyen

The Problem Definition and Refining

We wanted to address the problem of how publications on the internet influences stock prices. Since shareholders are keen on the success of companies they are invested in, we wanted to know if any online posts can influence their investments.

We decided to refine the problem by focusing on a single stock index: “S&P 500” and by looking into financial news articles that mention it in headlines.

Data Gathering and Cleaning

We searched for many datasets and settled on one that had financial news articles dating from 2010 to 2020 on Kaggle. The raw data had duplicated news articles since they only listed a single stock in a stock column and an article may mention multiple stocks. We removed these duplicates and grouped articles with the same date. We then filtered the data to only include mentions of S&P 500 or its aliases (S&P500, SP500, ^GSPC).

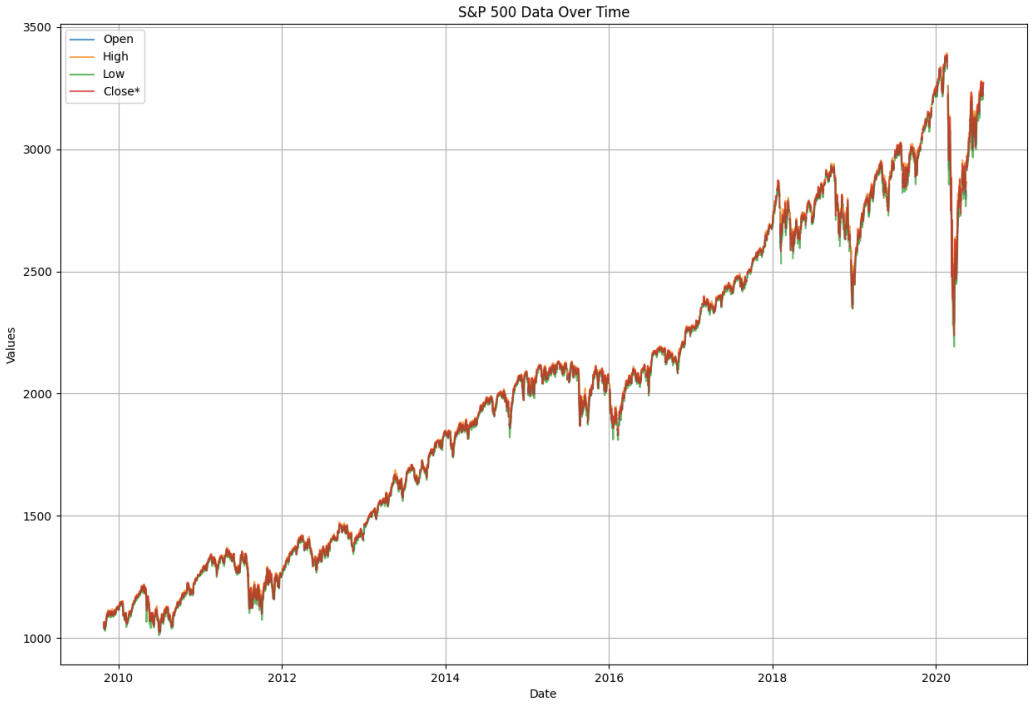
To retrieve the stock data, we scraped the historical stock data from 2010 to 2020 from Yahoo finance. We also included data of the S&P 500 stock prices shifted forward by a day so we can compare the price of that day and the prices of the day after.

We then combined the data from the news articles and the stock prices by syncing up their dates and exporting them into a single .csv file (master_data_set.csv).

Data Analysis

Visualizing the Data

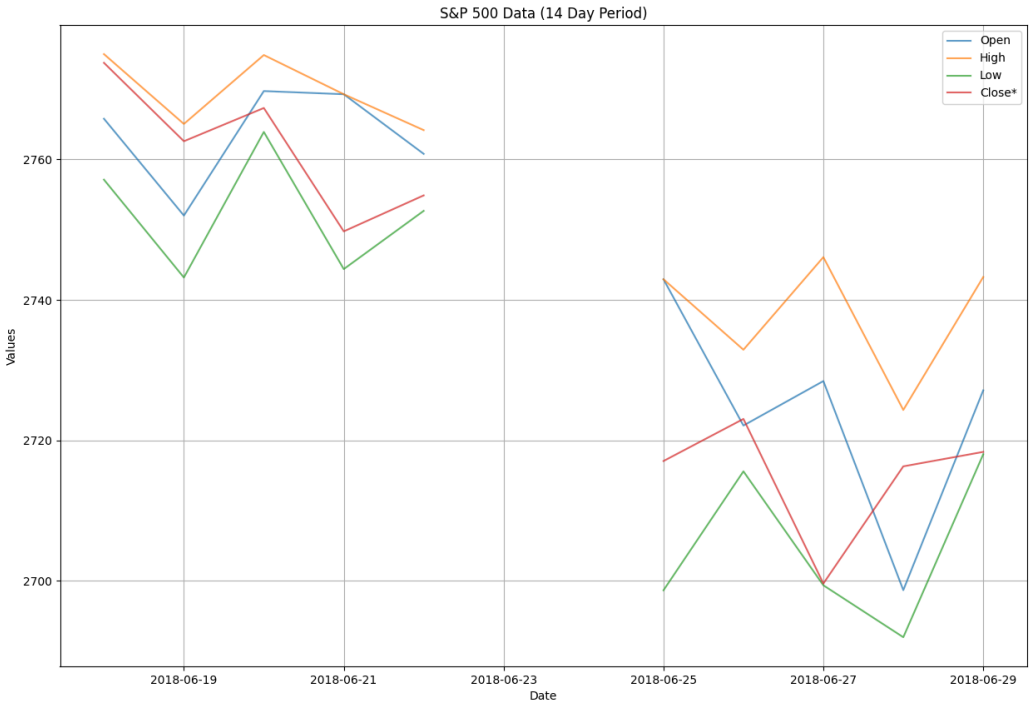
To better understand the data, what better way than graphing things out. Firstly, we plotted our base data of [Open, High, Low, Close] against time.



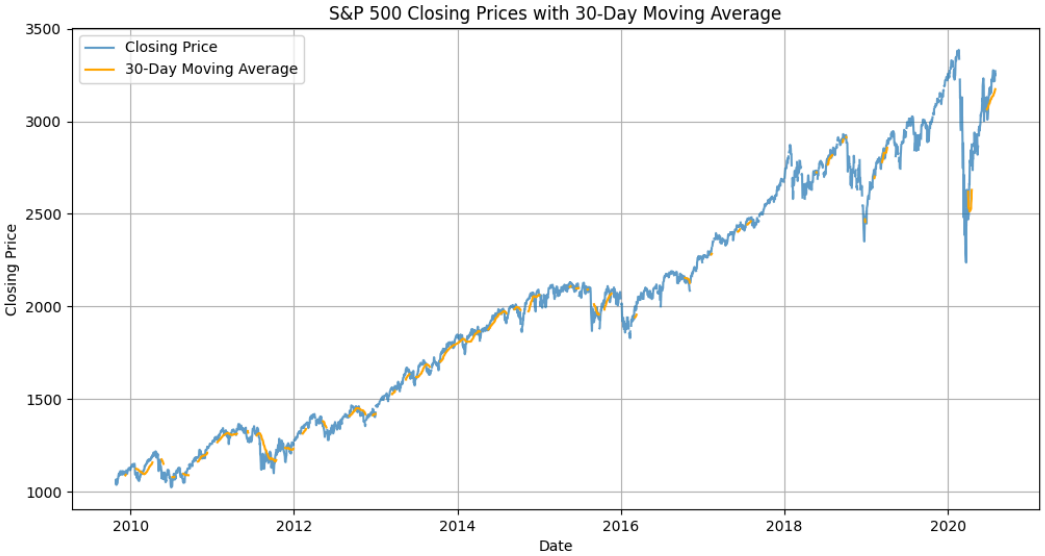
Open - The price that the stock opened on the day (9:30 AM EST)
High - The highest recorded price that the stock sold for within the day
Low - The lowest recorded price that the stock sold for within the day
Close - The last price that the stock sold at close (5:00 PM EST)

While the plot above is indeed very beautiful, it is difficult to visualize small details because there is so much data. We can take a snapshot of 3 months to get a better visualization of the data.

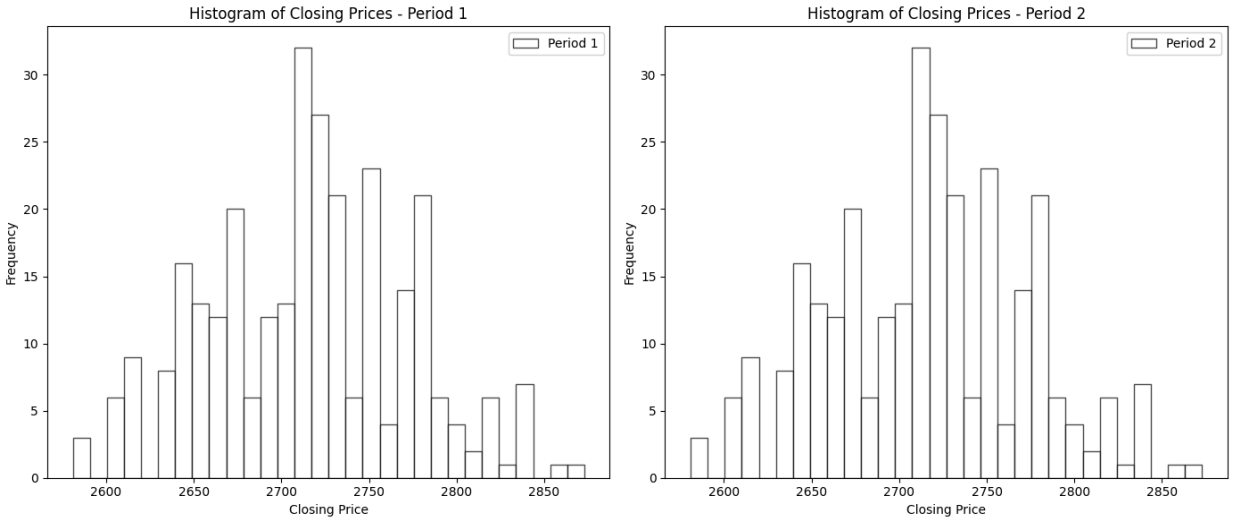

S&P 500 Data (3-Month Period)
Open
High
Low
Close*
Values
Date
2850
2800
2750
2700
2650
2600
2018-05-01
2018-05-15
2018-06-01
2018-06-15
2018-07-01
2018-07-15
2018-08-01


S&P 500 Data (14 Day Period)
Open
High
Low
Close*
Values
Date
2760
2740
2720
2700
2018-06-19
2018-06-21
2018-06-23
2018-06-25
2018-06-27
2018-06-29

Interestingly, we now observe gaps in the data in the 3 month plot. However, it is difficult to tell without taking an even smaller snapshot of 2 weeks where we note the gaps only occur during days of the weekend. This will be helpful to know in our further analysis, but also these gaps are expected since the stock market only operates during weekdays.



Sometimes stocks can be very volatile and no one likes seeing their net worth decrease at close. Because of this, we want to investigate the closing price parameter further and see if we can find any useful information.



As we explore our data further, we created a histogram and noticed that closing price is quite normal which is necessary to perform a T-Test. These charts let us visually inspect how the closing prices are spread out for each period.

The t-test looks at the average closing prices from two different periods to see if there's a significant difference between them. The p-value tells us if this difference is statistically meaningful. The normality test checks if the closing prices in each period follow a normal distribution. If the p-value is less than 0.05, it means the data isn't normally distributed. Performing the T-Test, our output is as follows:

T-statistic: -9.852659667117129, P-value: 1.7452485151021404e-21

Following this, we began our investigation by comparing closing prices of the year of covid (2020) and precovid (2019).

Hypothesis: Major economic events cause the average closing price of the S&P 500 to drop significantly.

Null Hypothesis (H0): The average closing prices of the S&P 500 before and after the economic event are not significantly different.

Alternative Hypothesis (H1): The average closing prices of the S&P 500 before and after the economic event are significantly different. The following output from our test:

T-statistic: -4.561026216150757, P-value: 5.890080181843442e-06

Since our p_value < alpha (0.05). We reject the null hypothesis: There is a significant difference in the average closing prices before and after the event.

Leading us to explore news headlines!

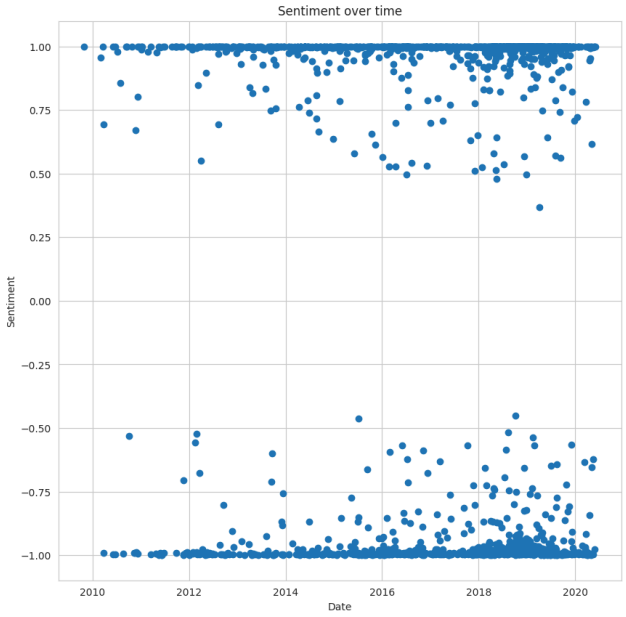
Techniques used to analyze the data:

To analyze the content of the news article headlines, we used Natural Language Processing (NLP) libraries to calculate the sentiment of the news. The sentiment is multiple numeric values that classify the headline as positive, neutral or negative. We used the TextBlob and Vader Python libraries and a model specifically trained on financial news from HuggingFace. We wanted to take the sentiment analysis from multiple models trained on different data to decide on a final sentiment result since there may be inaccuracy issues if just one model calculates the sentiment values.

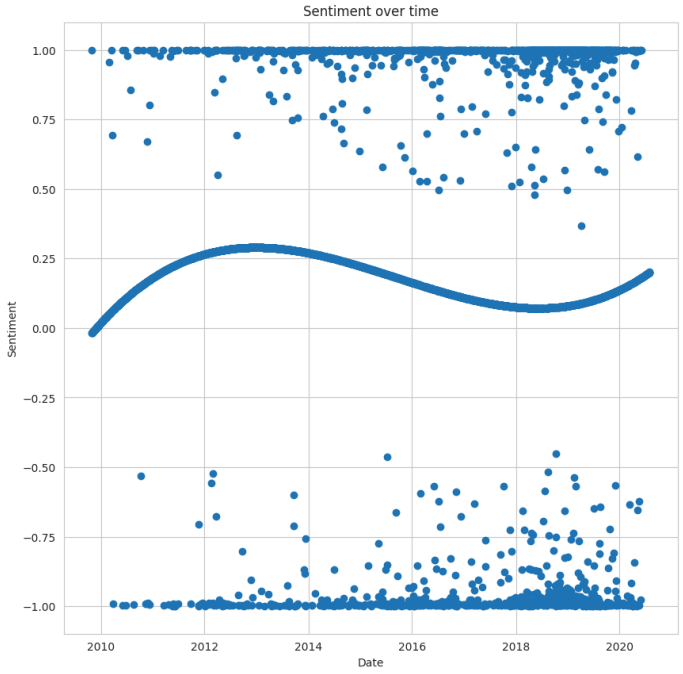
These output sentiment values were then averaged and used in predicting the next day’s stock price from the current day.

Training Data with Sentiment Value and Stock Information for 1-day prediction

1. Normalized sentiment values into [-1, 1]

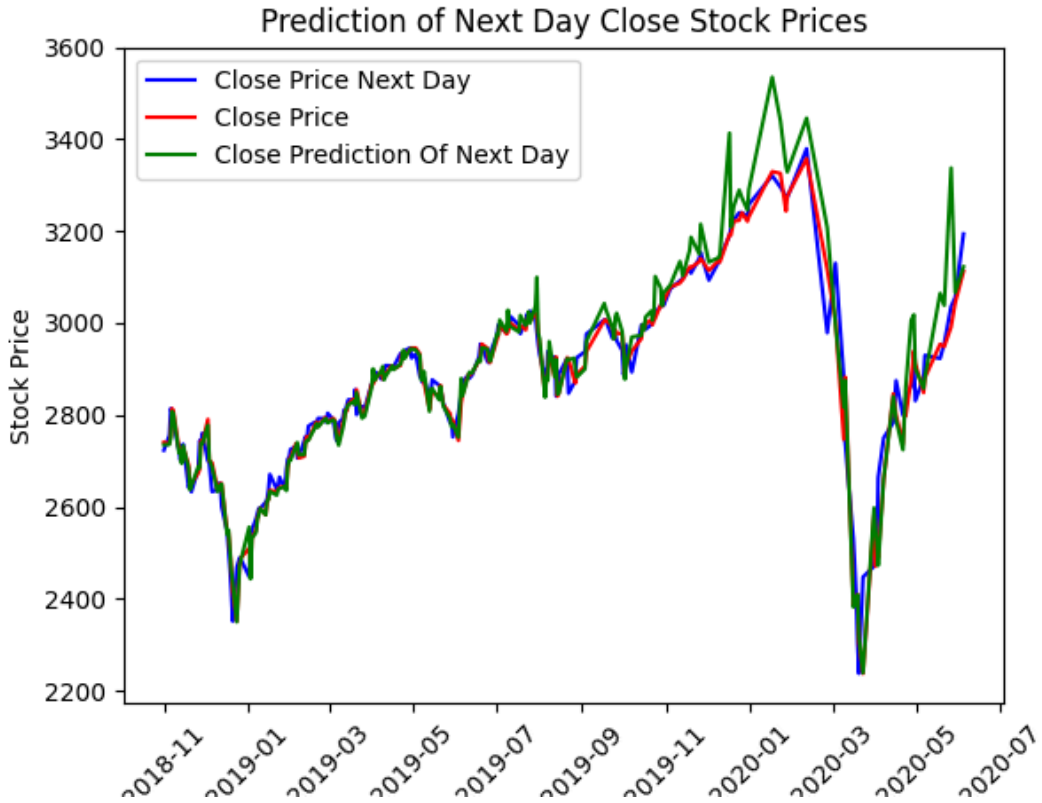


2. Interpolated sentiment values since we only have values at some specific location so the training data will be filled

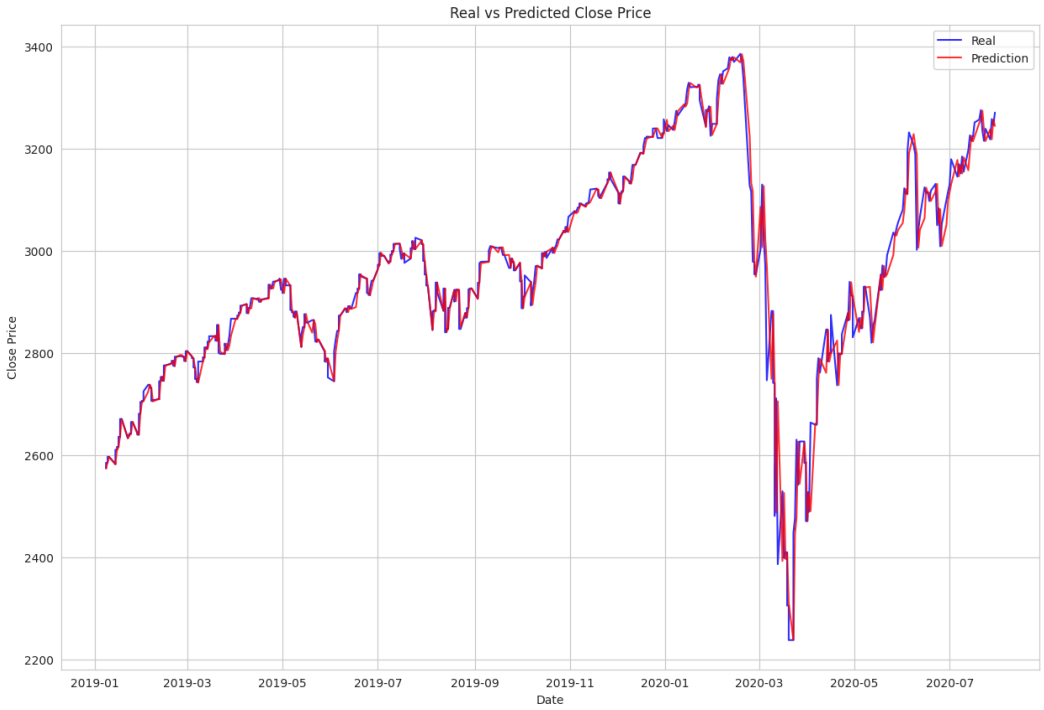


3. Duplicated closing stock price and shift it down by 1 so the training data can predict the next date avg. closing price
4. Split training and testing data 80-20 with no random shuffling because this data is following timeline

5. Make linear regression model based on train data
    a. Make 1st model with only previous closing price and sentiment values (no interpolation)
    b. Make 2nd model with basic stock information (open, closing, volume, high, low) and sentiment value with interpolation on missing data
6. Apply the model to test data and compare with observation value in test data
    a. First model:
        i. Average Differences = 36 USD
        ii. Mean Square Differences = 3309 USD
    b. Second model:
        i. Average Difference = 12.6 USD
        ii. Mean Square Difference = 1015 USD
7. Visualization results
    a. First Model:



    b. Second Model:



Conclusions

Results/findings/conclusions.

Our investigation shows that news headlines do have a significant impact on the S&P 500. Moving forward, we could build on more features such as congress trading or even features such as stock earnings announcements which can vastly improve our accuracy.

In our T-test, we reject the null hypothesis. There is a significant difference in the average closing prices before and after the event. This suggests that news headlines can have an effect on the S&P 500.

We also found that we cannot use one of our models, RandomForest, because it can only predict data in the range of training data which is not fit for stock prediction.

Limitations

We realized that the historical stock data does not have data on the weekends and news articles can be published on those days which results in missing and inaccurate values that we had to fill in ourselves. There was also a focus on regression rather than classification but sentiment values cannot accurately predict the amount of increase or decrease in stock prices alone. We may need other data to determine the magnitude of change based on a sentiment value.

Furthermore, we only try to predict the stock prices the day after the news articles are posted. Since the stock prices are fairly close together in a day, the accuracy of the prediction model may be inflated.

In retrospect, we should have added other social media data as features that our regressors could have used to predict the prices. That way, a more accurate change of stock price per day could be predicted.

Different companies and stocks may have different online presences as well. We should have picked the GameStop stock and used reddit data to predict how stock prices would fall or rise since it seems like they were largely influenced by Reddit a few years ago.

Project Experience Summary

Project Overview Accomplishment Statement:

Searched and filtered over 7000 financial news articles on Kaggle to look for trends in stock prices. Added Yahoo historical stock data with the dates of the news articles to predict stock prices. Utilized TextBlob, Vader and a HuggingFace model for sentiment analysis to predict trends in stock prices. Visualized the predicted data using Matplotlib for a better presentation.

Timothy:

Looked for a dataset on Kaggle for financial news articles for data analysis. Filtered 7000 financial news articles with some duplicates based on search criteria using Python to clean the data. Performed sentiment analysis on articles using Natural Language Processing libraries and

HuggingFace Models for machine learning. Created a basic Python script that uses machine learning to predict S&P500 close stock prices.

Lu:

Scraped data S&P 500 data from Yahoo Finance. Created master data through cleaning, truncated in complete data or irrelevant (1970-2009) because there was no news and merging news headlines with S&P 500 data. Analyzed sentiment values of news and interpolating missing data. Make prediction model using Linear Regression with stock information and sentiment values

Ernie:

Looked into creating plots to visualize our dataset and performed a null hypothesis test with closing price of stocks and major news. Initially scraped data for congress trading for coca-cola from quiver. Unfortunately, coca-cola data was lacking for news headlines, so we opted for s&p 500 which made it quite complicated to analyze.