Justin Jiang

Final Report

Project Abstract

There are instances where the price of a stock fluctuates depending on the sentiment of their investors, also known as market sentiment. The goal of this project is to create forecasting models where it takes market sentiment into account. The approach of this project was to find the average of sentiment scores in tweets to capture market sentiment for that day. SARIMA models were built for each chosen company to serve as a base when constructing SARIMA models with market sentiment included. SARIMA models are chosen since it is effective in forecasting time series and exogenous variables like market sentiment can be easily included.

The project's finding shows that the market sentiment had barely impacted the forecast of the constructed SARIMA model. The forecasts from the SARIMA with market sentiment were similar to that of the forecast of the base SARIMA model. This led to the SARIMA model of the market sentiment having a similar trend with the base SARIMA model.

There are limitations presented in this project. The data is limited to one year of tweets and stocks. Evaluating any monthly patterns have to assume to be true for all years. Another limitation is that the project limits itself to the use of simpler forecasting models being SARIMA. The use of complex models like recurrent neural networks can be explored in the future to test the effectiveness of market sentiment. Furthermore, not all days in the forecasting model don't contain an equal amount of tweets per day for each company. This would mean that one tweet can represent market sentiment for that day. These tweets were kept because they seem valuable.

Accomplishments

The major goal of our project was to create a forecasting model that uses the market to effectively capture a short term forecast. Once achieved, the goal is to see if market sentiment would have an impact on forecasting models for predicting stock prices.

Specific Objectives

All of aim 1 was completed which consisted of two sub aims. Sub aim 1 was focused on preprocessing text and feature engineering suited for sentiment analysis. Sub aim 2 focuses on obtaining sentiment analysis and later using the scores from each tweet to calculate the market sentiment of that day.

Aim 2 was to create a forecasting model on the stocks given to us. sub aim 1 was to create a forecasting model like ARIMA as base so that exogenous variables, specifically sentiment scores can later be used to help build on that model for predicting stock prices. Although we accomplish most of aim 2, more time could be needed to further investigate possible models to build models that use sentiment scores for better predictions.

Major Activities

Data visualization and sampling: Investigated to see how the number of tweets were distributed for each company. The dataset contained a tremendous amount of tweets for the duration of the one year (Sep 2021-Sep 2022) for Tesla (TSLA). Other companies like Amazon (AMZN), Microsoft (MSFT), and Procter & Gamble (PG) have the same amount of tweets. The stock prices of each company were forecasted as well. There were companies that showed some trends while others were volatile. The companies were chosen based on the movement of their stock price and the number of tweets a company has. TSLA was chosen due to the numerous tweets and wanted to investigate with prices being volatile. MSFT was chosen due to a sufficient number of tweets and that the prices are non-volatile. These will be used to evaluate the performance of forecasting models.

Subaim 1: To manage with the text data, the tweets, the utilization of the nltk library was used to deal with the preprocessing of the text data. The preprocessing steps were to first, remove hashtags, links (http), and @ and any preceding word after that. Due to the nature of tweets containing emojis, they were kept. The changed tweets were later tokenized. Common stop words were removed, lemmatization was used to remove words that are variations of a word, and punctuations were removed. The tokenized tweets became its own feature for sentiment analysis.

Subaim 2: After preprocessing the tweets, I used the emoji library to convert the emojis in lemmatized tweets into words so that the Textblob library can properly assess the sentiment of each tweet. Each lemmatized tweet with the emojis converted were given a score between -1 to 1 based on how positive/negative each tweet was. After identifying the score, the average sentiment was calculated based on the company associated with each tweet and the specific date. This is so that the market sentiment towards a company can be acquired for that day. I implemented a labeling system depending on the sentiment score of a tweet to help with finding the average sentiment. This labeling will help be used to assist in a weighted average approach toward finding average sentiment. The majority class will apply more weight to the scores belonging to that class. After finding the average sentiment, I merged the average sentiment with the dataset containing the company's stock prices.

Aim 2:

Subaim 1: For both TSLA and MSFT, I decomposed their stock prices to analyze their trend, seasonality, and residuals. After determining any patterns, differencing was performed to understand if the prices over time were stationary. ACF and PACF plots were built to determine potential parameters in case auto regressive and moving average is needed. SARIMAs model was built with investigated parameters for each company. Forecasted the next five days of the stock prices for each company to compare with the actual prices to determine performance of the model. RSME (root squared mean error) was used to test the accuracy of the forecasts, and if the SARIMA models captured the decomposition of each time series of the stock prices.

Subaim 2: After determining suitable SARIMA models as a base, SARIMA models were performed with the same models again with an exogenous variable (average sentiment) instead. Forecasts of the next 5 days were made based on the most recent sentiment score. Forecasts were made again, but based on the 5 most recent sentiment scores instead to see if past market sentiment had an impact and not recency. RSME was the same metric used to test the accuracy of the forecast and distance metric to see the similarity between the forecasted results from the base SARIMA model and SARIMA with the exogenous variable.

Significant Findings

For the sentiment analysis, there weren't many tweets containing severe feelings for the market (Fig 1). The scores might not have a very significant impact in price change due to most market sentiment leaning towards a bullish or neutral attitude.

Fig 1

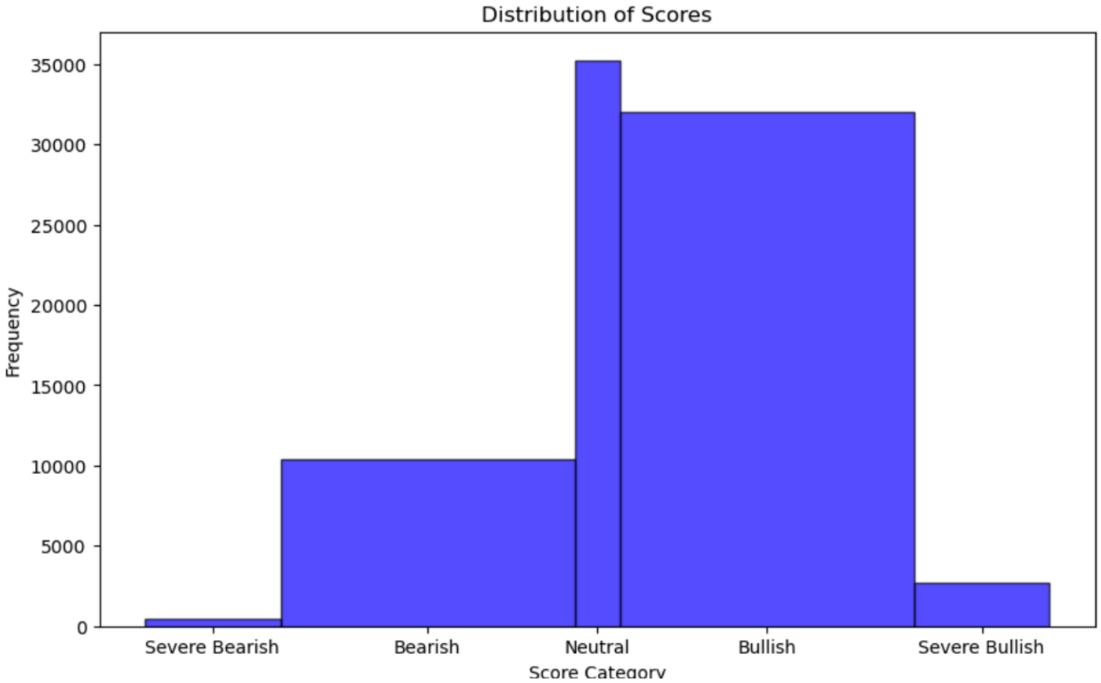


Fig 2.1

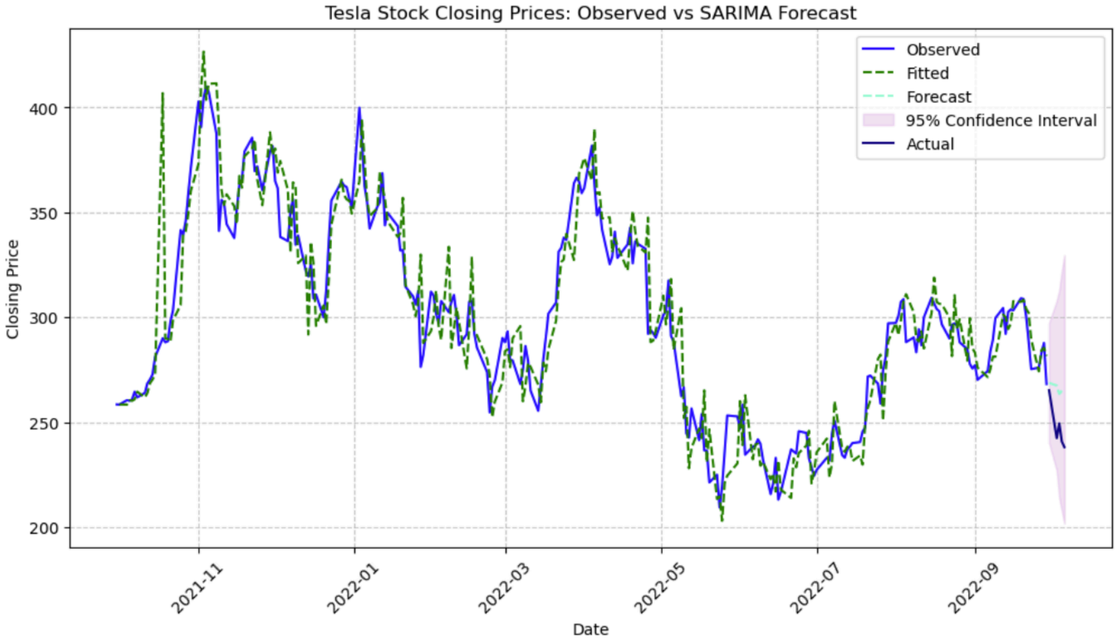
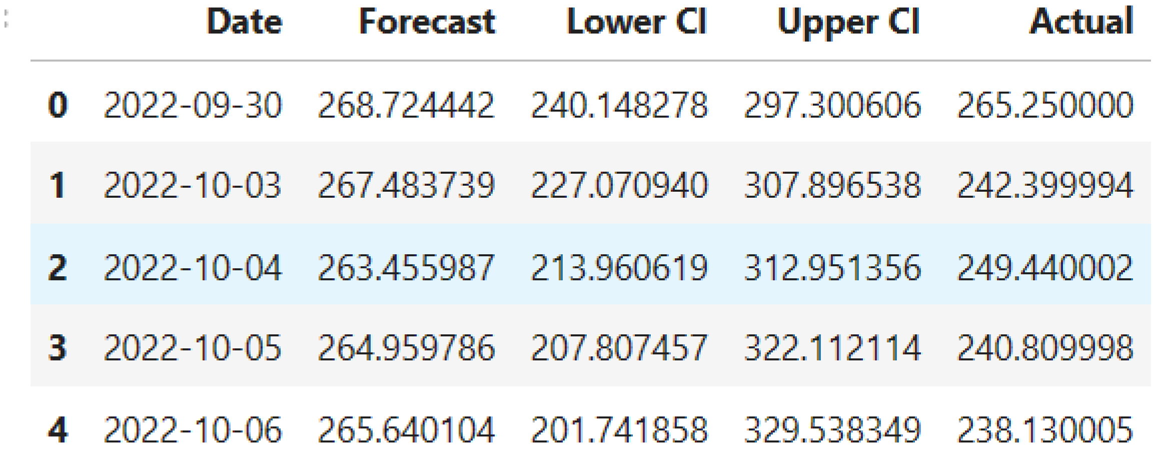


Fig 2.2



	Date	Forecast	Lower CI	Upper CI	Actual
0	2022-09-30	268.724442	240.148278	297.300606	265.250000
1	2022-10-03	267.483739	227.070940	307.896538	242.399994
2	2022-10-04	263.455987	213.960619	312.951356	249.440002
3	2022-10-05	264.959786	207.807457	322.112114	240.809998
4	2022-10-06	265.640104	201.741858	329.538349	238.130005

After investigating that the time series is not stationary, looking at the ACF and PACF plot, and testing the seasonal order's parameter. The parameters were order is (0,1,0) and seasonal order is (1,1,0,12). Fig 21 displays the observed values of Tesla's closing price compared to the SARIMA fitted and forecasted price. The RSME between the observed price and the fitted price is 16.37. The RSME between the future observed price and the forecasted price in Fig 2.2 is 20.869.

Fig 2.3

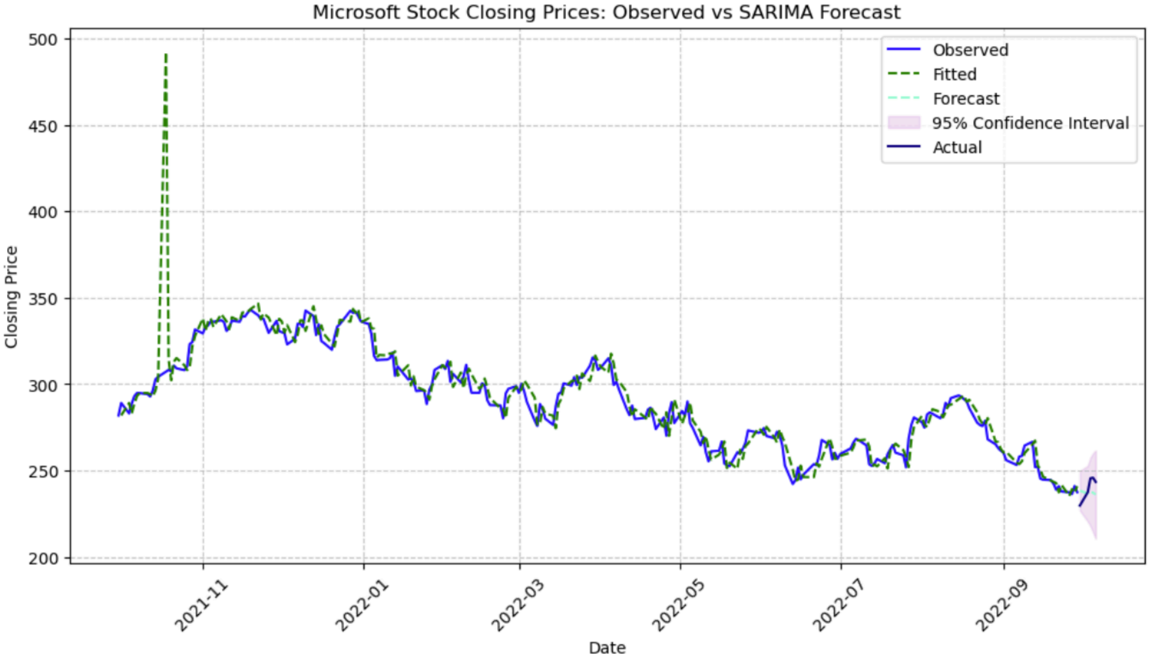
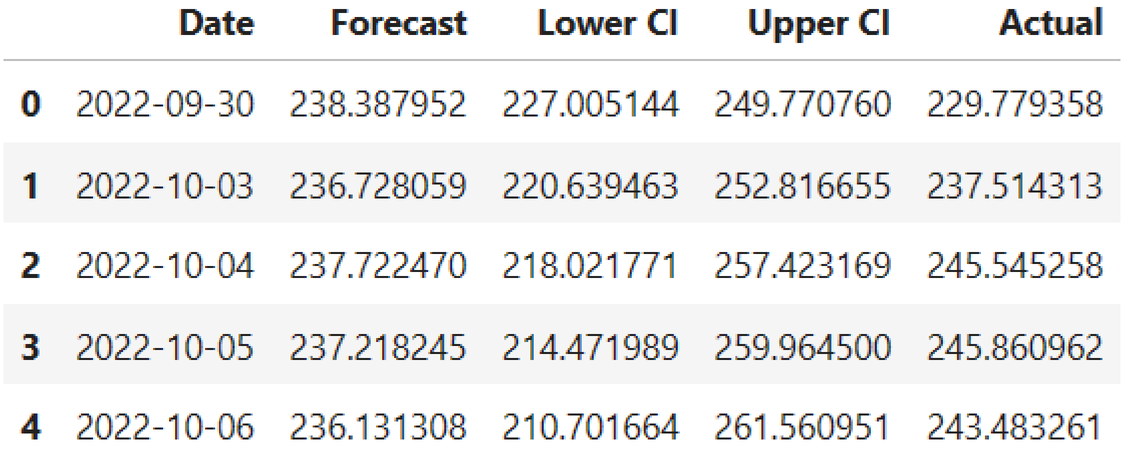


Fig 2.4

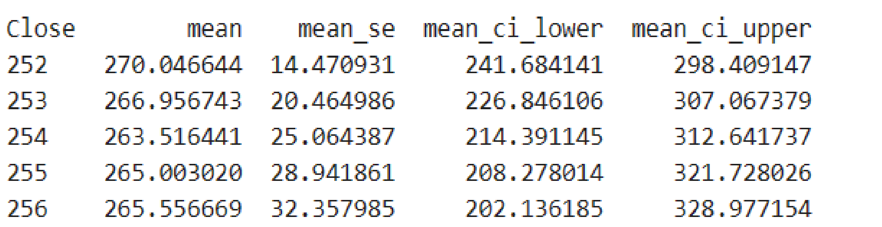


	Date	Forecast	Lower CI	Upper CI	Actual
0	2022-09-30	238.387952	227.005144	249.770760	229.779358
1	2022-10-03	236.728059	220.639463	252.816655	237.514313
2	2022-10-04	237.722470	218.021771	257.423169	245.545258
3	2022-10-05	237.218245	214.471989	259.964500	245.860962
4	2022-10-06	236.131308	210.701664	261.560951	243.483261

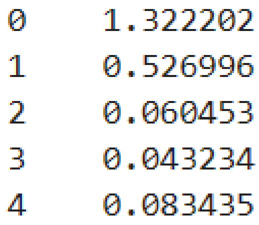
After investigating that the time series is not stationary, looking at the ACF and PACF plot, and testing the seasonal order's parameter. The parameters were order is (0,1,0) and seasonal order is (1,1,1,12). Fig 2.3 displays the observed values of Microsoft's closing price compared to the SARIMA fitted and forecasted price. The RSME between the observed price and the fitted price is 77.77. The RSME between the future observed price and the forecasted price in Fig 2.4 is 7.275.

Fig 3.1.1

Fig 3.1.2



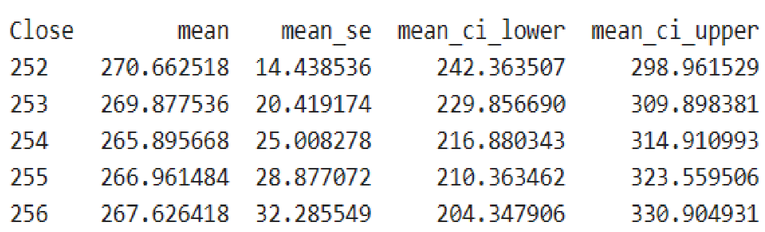
```
Close         mean     mean_se  mean_ci_lower  mean_ci_upper
252     270.046644   14.470931     241.684141     298.409147
253     266.956743   20.464986     226.846106     307.067379
254     263.516441   25.064387     214.391145     312.641737
255     265.003020   28.941861     208.278014     321.728026
256     265.556669   32.357985     202.136185     328.977154
```



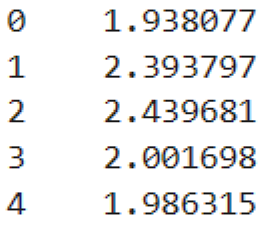
```
0     1.322202
1     0.526996
2     0.060453
3     0.043234
4     0.083435
```

Fig 3.2.1

Fig 3.2.2



```
Close         mean     mean_se  mean_ci_lower  mean_ci_upper
252     270.662518   14.438536     242.363507     298.961529
253     269.877536   20.419174     229.856690     309.898381
254     265.895668   25.008278     216.880343     314.910993
255     266.961484   28.877072     210.363462     323.559506
256     267.626418   32.285549     204.347906     330.904931
```



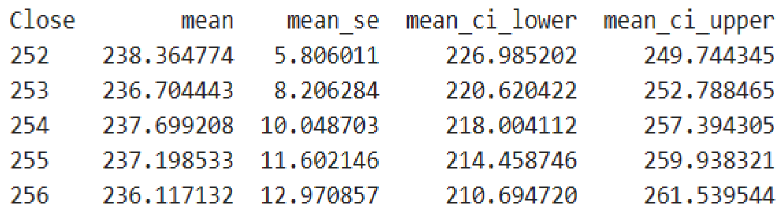
```
0     1.938077
1     2.393797
2     2.439681
3     2.001698
4     1.986315
```

Fig 3.1.1 shows the forecast of Tesla's stock price in the next 5 days based on the most recent sentiment score. Fig 3.1.2 shows how similar the forecast is with Tesla's forecasted prices from the base SARIMA model. Fig 3.2.1 and Fig 3.2.2 is the same thing but the forecast is based on the 5 most recent sentiment scores.

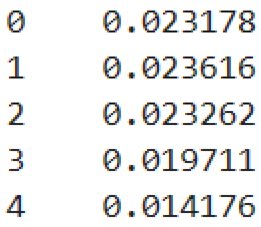
The RSME from the forecast based on the most recent market sentiment is 20.79. While, the RSME from the 5 most recent forecast is 22.8.

Fig 3.3.1

Fig 3.3.2



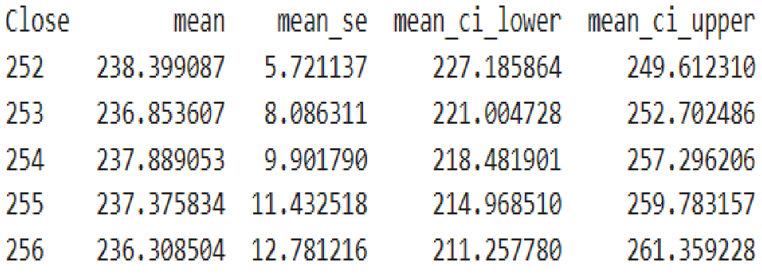
```
Close         mean     mean_se  mean_ci_lower  mean_ci_upper
252     238.364774    5.806011     226.985202     249.744345
253     236.704443    8.206284     220.620422     252.788465
254     237.699208   10.048703     218.004112     257.394305
255     237.198533   11.602146     214.458746     259.938321
256     236.117132   12.970857     210.694720     261.539544
```



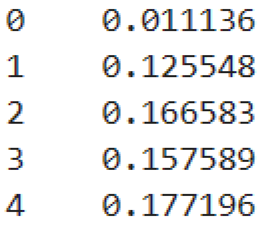
```
0     0.023178
1     0.023616
2     0.023262
3     0.019711
4     0.014176
```

Fig 3.4.1

Fig 3.4.2



```
Close         mean     mean_se  mean_ci_lower  mean_ci_upper
252     238.399087    5.721137     227.185864     249.612310
253     236.853607    8.086311     221.004728     252.702486
254     237.889053    9.901790     218.481901     257.296206
255     237.375834   11.432518     214.968510     259.783157
256     236.308504   12.781216     211.257780     261.359228
```



```
0     0.011136
1     0.125548
2     0.166583
3     0.157589
4     0.177196
```

Fig 3.3.1 shows the forecast of Microsoft's stock price in the next 5 days based on the most recent sentiment score. Fig 3.3.2 shows how similar the forecast is with

Microsoft's forecasted prices from the base SARIMA model. Fig 3.4.1 and Fig 3.4.2 is the same thing but the forecast is based on the 5 most recent sentiment scores.
The RSME from the forecast based on the most recent market sentiment is 7.28. While, the RSME from the 5 most recent forecast is 7.16.

Based on the distance measured between the base SARIMA forecast and the SARIMA forecast with sentiment scores, it was concluded that the market sentiment did not have a significant impact on effectively predicting stock prices with the created forecasting models. This also means the sentiment scores had any impact on the trend of the SARIMA since the forecasted results are similar with the base SARIMA. The RSME between the base SARIMA and SARIMA with sentiment scores were similar as the difference between them were close (between 0.1 to 2). Furthermore, when finding how similar the forecasted results from the SARIMA were compared to the forecasted results from the SARIMA with sentiment scores, they were shown to be very similar since the distance between them is less than 2.5. If the RSME were smaller or bigger, it would have been concluded that market sentiment did have an effect. The lack of impact might have occurred due to market sentiment not having any weight into the forecasting model.

Other Accomplishments
The use of ARIMA was used but it provided irrelevant results as it will be explained in the problems.

Changes/Problems

Changes in Approach

1. For preprocessing, I changed my approach where I kept emojis to be converted to words instead of erasing them since a lot of the tweets consisted of emojis.
2. Changed the sampling approach to the project where I just compared the forecasting model of two companies. This was due to a change in testing the effects of market sentiment on volatile and non volatile data.
3. Didn't use the measure of R^2 as a metric because in hindsight it would not tell much information about the similarity of the forecasted values later on. RSME was sufficient in telling both performance and similarity of a model.
4. Didn't use ARIMA because the forecasts it provided were constant.
5. Forecasted 5 days instead of 7 days because that's the availability of the stock market within one week.
6. Didn't use regression techniques to check the direction of the forecasted prices since checking the trend of the forecasted price wasn't really needed.

Problems or Delays Experienced and Corrective Actions Taken

During the process of creating an ARIMA model for Tesla, the forecast produced for the next 5 days were the same. Even when plotting the forecast, it was very noticeable that the forecast was the same. To address this, the ARIMA model was changed to a SARIMA where seasonality was used because of the inclusion of seasonal patterns being used.

Another issue was how to use exogenous variables (market sentiment) into the forecasting model without future market sentiment available. To tackle this problem, I used the most recent market sentiment as a placeholder for the unavailable market sentiment. This uses the most recent sentiment score. For the 5 most recent sentiment scores, I used rolling average where it took the 5 most recent sentiment scores as a substitute for the future market sentiment. When dealing with sarcasm, I used a weighted average to find the market sentiment, where the scores belonging to the majority class had more weight.

Impact

Impact on the Domain

These findings provide researchers or traders information on the effectiveness of using exogenous variables like market sentiment on predicting stock prices. It arrives at the conclusion that the change of stock prices doesn't really reflect the market sentiment. Thus, it is not possible to predict stocks in the short term by just using market sentiment. Predicting stock prices proves to be a complex task as the prices are dynamically changing. Although this is well known among traders and researchers, this project displays predicting stock prices isn't so simple with common factors like market sentiment taken into account. This project also proves that more complex models such as recurrent neural networks or additional features will be needed to predict stock prices as changes in prices don't reflect how investors feel and that it is not easy to capture patterns of stock prices.

Impact on the Individual

I have learned a lot for the duration of the project. The inspiration for this project came from when my friend showed me a paper discussing models used to predict stock prices. Forecasting was one of them. Also, I have been intrigued with the use of features outside of a time series to predict stock prices. Coming into this project, I was a novice in modeling with time series data and the only experience I had with time series was just using exponential smoothing. I learned how to tackle text data, especially with data containing multiple emojis and sarcasm. This project also taught and allowed me the tools used to work with modeling with time series and what models can be used with features outside of the base model. This project taught me that looking at different time series can be complex, and it challenges me to explore more complex models to effectively predict them. In the future, I am inspired to work in industries like the finance industry and understanding how to work with time series helps me understand the complexity and methodology when approaching a time series like sticks. This is especially true in the field of finance where events have an impact on changes in investments.