Predicting Tesla Stock Price Changes Using Elon Tweets

CMPT 353
Fall 2023
HiFen Kong and David Choi

Introduction

In this project, we investigate the correlation between Elon Musk's Twitter tweets and Tesla's stock TSLA. Using a dataset made up of Elon Musk's Tweets, TSLA stock data, and S&P 500 stock data, we will analyze the correlation between Musk's Tweets and TSLA stock, compare TSLA trends with the overall market trends, the historical changes in the datasets, as well as a possible machine learning models that can predict fluctuations in the TSLA stock price using features of Musk's Twitter tweets. We hope these findings can help determine the influence of Musk's tweets on investor confidence in Tesla and its implications on the stock market.

ETL

Elon Tweet Data:

To handle the Elon Musk tweet data, we read the Elon Musk tweet dataset into a PySpark data frame. Before going through any analysis, we had to understand and clean up the data so that only the relevant tweets remained.

When going through the dataset, we realized that the timestamps of the tweets were in the UTC time zone, which meant that we had to compare it with the US market times in the UTC time zone. We started by shifting the tweets made after the market closed at 20:00 UTC. Then, we moved these tweets to the following day as that would be the first day the market could be affected by the tweet. Subsequently, we removed the time from the date as it was no longer relevant. After this, tweets made on holidays or weekends were shifted to the next open market date so they would reflect accurate trading actions. We first prioritized transforming the tweets made after market closing time, as this could potentially push the date to a holiday or weekend.

Next, we filtered out the non-English tweets from the data. We had to filter the non-English tweets from the data after the date-time transformations, as we wanted to compare the non-filtered with filtered data and the date change was necessary for both datasets. Afterward, we filtered out the tweets with fewer retweets than the median count. We did this because we only wanted tweets that had reached a larger audience as we believed that would have a higher likelihood of reaching investors, affecting the market price of the TSLA. We used the median instead of the mean of the retweets as it is not as heavily skewed by outliers.

Finally, we converted the text of the tweets to lowercase and filtered out the ones that did not contain the keyword 'tesla'. Moreover, we dropped some of the columns that were not necessary in our analysis, keeping only the relevant ones, before writing out the data into a set of filtered and non-filtered .csv files.

Tesla Stock Data and S&P 500 Stock Data:

For the Tesla and S&P 500 stock data, we read the data into a Pandas data frame before dropping the irrelevant columns, keeping columns needed for analysis. The change percent is the percent change between the closing price of the stock and the opening price of the stock for that day. Then, we removed the rows with the invalid data and wrote it out into two separate .csv files.

Joining The Data:

The Twitter data has a sentiment column that is generated by this model. The column contains the sentiment type and its sentiment score. The sentiment type is a label that describes the sentiment as positive, negative or neutral. Additionally, the sentiment score is a score that measures the intensity of the sentiment type with 0 being low and 1 being high. We initially separated those two values into separate columns, which allowed us to group the tweet data into separate data frames by sentiment types. Then, we joined the positive, neutral, and negative sentiment-filtered and unfiltered tweet data with the S&P 500 Stock data and the TSLA stock data by their dates and wrote it out into multiple .csv files. Additionally, we merged the S&P 500 stock data with the TSLA stock data on their dates as well and saved the combined dataset into a single .csv file.

Correlation Analysis

We first compare the correlation between the tweet data and the TSLA stock to see the effect of the tweets on the stock price. We found the correlation for both the filtered and unfiltered tweet data with the TSLA to be very small. Even with the small correlation, the filtered tweets had a better correlation with the TSLA stock than the nonfiltered tweet data for most sentiment types. This makes sense as most general tweets unrelated to Tesla would not have a significant influence on investors. However, since the correlations are very small, we can see that the linear relationship is very weak. Fig. 1 shows the correlation between the filtered sentiment score and the change percent. The values of those correlations are printed below.

```
Positive Filtered Change Percent:  0.11127646415388137
Negative Filtered Change Percent:  0.04204682394122923
Neutral Filtered Change Percent:  -0.07071663369744226
```

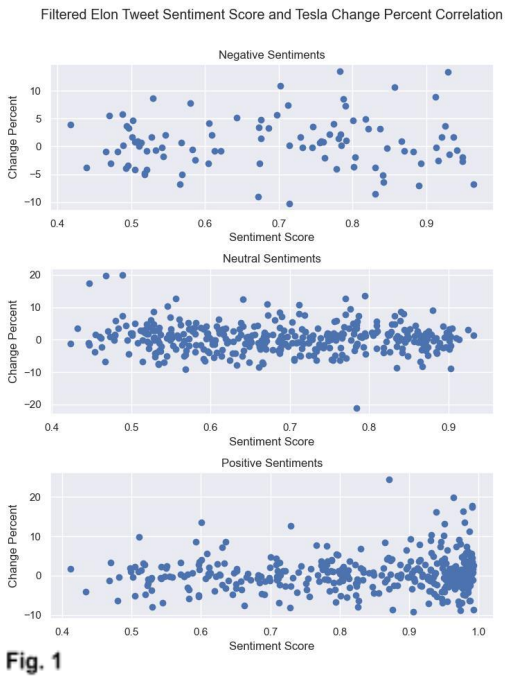

Fig. 1

We also compared the correlation between the tweet data and the TSLA stock to see the effect of the tweets on the volume of stocks being traded. Similarly, as with the correlation between tweet sentiment scores and change percent, the correlation for both filtered tweets and unfiltered tweets here is also very small. Again even with the small correlation, the filtered tweets had a better correlation with the TSLA stock than the nonfiltered tweet data for most sentiment types and with the small correlations, we can see that the linear relationship is very weak. Fig. 2 shows the correlation between the filtered sentiment score and the volume. The values of those correlations are printed below.

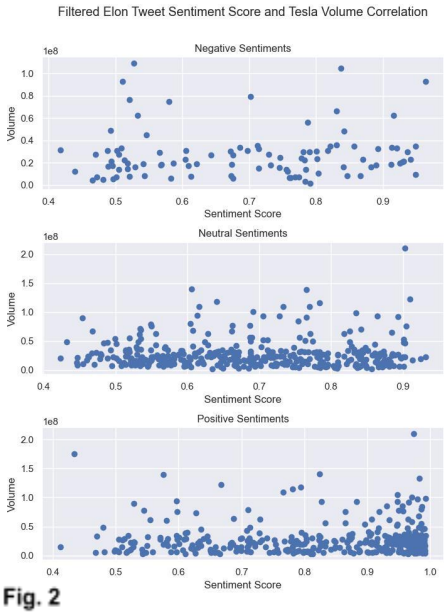

Fig. 2

```
Positive Filtered Volume: -0.006183891640038872
Negative Filtered Volume:  0.05239380387589901
Neutral Filtered Volume:   0.03517908359933082
```

Historical Analysis

Our goal was to do a historical analysis on the negative, neutral and positive filtered tweet data and TSLA stock data to see if there were any major differences in sentiment score or change percent in any year. If there were any years with any major differences in mean we hoped to see similar differences in other datasets in the same year. To do this we separated each of the datasets into separate data frames by the year of the data points. In doing so we noticed that in the years between 2010 - 2016, there were empty data frames for the tweet data of most sentiment types, this was due to the lack of any tweets related to Tesla at that time. We then performed the ANOVA test on each set of data frames. Of the 4 ANOVA tests we ran, only the TSLA had a p-value of significance (`0.016649`). The others all failed to reject the null hypothesis and thus there was no major difference in sentiment score for each sentiment type throughout the years. For the TSLA data since it could reject the null hypothesis, we did post hoc analysis using the Tukey Test to see which years had a significant difference between their mean change percentages. What we found was that only the years 2020 and 2022 had significant differences, with 2020 having a higher positive mean change percent compared to most years and 2022 having a higher negative mean change

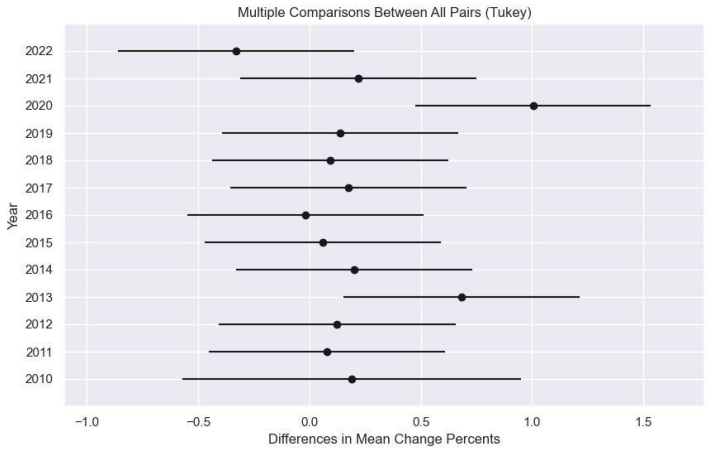

Fig. 3

percent compared to most years. Since there was little deviation in the sentiment score of any of the filtered tweet data of any sentiment type it appears the datasets did not share any distinct patterns in the years they had major differences.

Machine Learning Model

For our machine learning, our goal was to use a regression model to determine the percent change of the TSLA stock price using features of Elon's Tweets. The features we used were the sentiment score, sentiment type, and retweet count. The sentiment type is not numerical data so we need to convert it to a number so it can be used as a feature in a regression model. We change the value to -1 for negative, 0 for neutral and 1 for positive sentiments, we can do this because the sentiment types were already ordinal data. Next, we created a pipeline, first by using a standard scaler that centers the values around the mean with a unit standard deviation. Then, we used the RandomForestRegressor model in the pipeline and got the training and validation score. We did the same again but with the KNeighborsRegressor model. Then, again with a VotingRegressor model which included the KNeighborsRegressor, RandomForestRegressor, and the GradientBoostingRegressor models. Looking at the score of the models we can see that the models do an extremely poor job of predicting the percent change of the TSLA using the features of Elon's tweets. This is expected as the correlation between these datasets seems to be very low. The best model with the least inaccurate predictions was the RandomForestRegressor, shown below.

```
RandomForestRegressor Scores
Training score:  0.07031898108399615
Validation score:  -0.008825964578065282
```

Conclusion

Our findings in this project show that there may be some small correlation between Musk's tweets and the change in price of the TSLA. Our filtering of Musk's tweets led to better correlations but only increased the correlation by a small amount. We also did a historical analysis to see if there were any major differences in sentiment scores or change percent in any year. We failed to find any pattern between the yearly differences in mean sentiment scores and the change percent of the TSLA stock. We then used regression models to predict the change percent in the TSLA stock price with features of the Tweet data. As expected from the low correlation between the datasets, the validation scores of these models were all extremely low. In conclusion, according to our analysis, while there may be some small relationship between Musk's tweets and the TSLA, there is no major linear correlation between the filtered tweets and

the change percent and the volume of stock traded. Furthermore, we can conclude that the tweets that we filtered do not have a large influence on investor conscience and the TSLA stock.

Limitations

A limitation we had was the filtering of the Twitter dataset. If we had more time we would like to find a better way to filter by keywords so that we can filter with more keywords than just “tesla” which would probably lead to better results from our filtering. Additionally, we would have looked into filtering the tweets on different features like retweet count, like count, etc.

Another limitation we had was the sentiment data collected in the Twitter dataset. We would have more flexibility with the sentiments if all the sentiment types and score for each type was recorded for every tweet instead of just the sentiment with the highest score. In retrospect, we probably could have looked into calculating those sentiment scores ourselves.

Project Experience

HiFen Kong

- Used PySpark to significantly improve the performance and runtime of the ETL program
- Made use of the Anova and Tukey tests to complete a historical analysis of our data
- Filtered and transformed the datasets to make them more suitable for our analysis

David Choi

- Created an algorithm that converts the dates of the invalid tweets to the next available open market times.
- Modified machine learning regressors and their parameters to increase the accuracy of regression scores
- Used matplotlib and sealion libraries to create informative figures to better communicate our findings

The Data

The datasets we used for this project were sourced from Kaggle.com an online AI and machine learning community that provides free datasets.

- **Elon Tweet Data:** Elon Musk Tweet Dataset
- **SP500 Stock Data**: S&P 500 Stock Dataset
- **Tesla Stock Data**: Tesla Stock Dataset