Tesla (TSLA) Stock Prediction

Using Sentiment Analysis of Tweets

By

Vanessa Tang

Akshay Punwatkar

Fintech 540

Machine Learning for Fintech

Duke University

(November 2020)

1. Introduction

In recent years, social media platforms such as Twitter and Facebook, have been increasingly reflecting and influencing complex systems, one of which is the financial market.[3] The extent of this influence is yet to be quantified. This project aims to assess the ability of social media information to predict stock market prices. The scope of this project focuses on Tesla, an electric vehicle and clean energy company known not only for its electric cars but also for its CEO, Elon Musk. Musk is an outspoken persona, seemingly more active on social media than any other CEO. Whether he's talking about Tesla's newest creation, discussing politics, or critiquing COVID-19, Musk seems to be a big voice on Twitter, becoming the first of what some may call the "CEO influencer." Musk has over 40 million Twitter followers, while Tesla has a mere 6.5 million, emphasizing Musk's popularity as a source of entertainment for many. How does this Twitter activity affect Tesla's stock price?

The goal of this project is to delve deeper into how Tesla and Elon Musk tweets affect the Tesla stock price. To do so, over the course of a few weeks, we gathered tweets from Tesla and Elon Musk in addition to tweets about Tesla and Elon Musk and performed sentiment analysis to predict the Tesla stock price.

In the process of forecasting the *opening* price of Tesla (TSLA) for the next day, we developed several predictive models utilizing the average sentiment value across tweets, and volume of tweets from the past days along with the *opening* and *closing* price of the previous day. However, our analysis showed an absence of any strong correlation between the sentiment or the volume with the stock price. Moreover, further analysing the movement of Tesla (TSLA) overtime via online literature also pointed towards the unpredictability of the stock, which some sources attributed to over-evaluation of Tesla[1][2] and stock inflation by the non-traders as a result of news media and handy access to trading via platforms like Robinhood.

Tweets, especially by Elon Musk[3], have previously affected Tesla's stock price, but the change in price may not exclusively be attributed to the sentiment of the tweets regarding Tesla. It appears the effect is more likely due to the validity or the facts represented in the tweets. Future analyses could perform more natural language processing techniques to analyze the content of these tweets to predict Tesla's stock price.

2. Data

2.1. Data Streaming

2.1.1. Yahoo Finance

The stock data was collected using Yahoo Finance API via Rapid API. We were able to stream the data for the past 10 years. The data consisted of the open, high, low, close and adjusted close price.

date	open	high	low	close	adjclose
2010-11-16	6.20	6.28	5.68	5.93	5.93
2010-11-17	6.04	6.15	5.72	5.90	5.90
2010-11-18	6.13	6.15	5.78	5.98	5.98
2010-11-19	6.03	6.27	5.94	6.20	6.20
2010-11-22	6.31	6.69	6.30	6.68	6.68

date	open	high	low	close	adjclose
2020-11-10	420.09	420.09	396.03	410.36	410.36
2020-11-11	416.45	418.70	410.58	417.13	417.13
2020-11-12	415.05	423.00	409.52	411.76	411.76
2020-11-13	410.85	412.53	401.66	408.50	408.50
2020-11-16	408.93	412.44	404.09	410.39	410.39

2.1.2. Twitter

For our analysis, we used the social media platform Twitter to gather tweets by Tesla and Elon Musk, aimed at Tesla and Elon Musk, and tweets about Tesla and Elon Musk. Tweets were pulled according to the following schema to generate three main categories:

1. Tweeted by Tesla (@Tesla)
2. Tweeted by Elon Musk (@elonmusk)
3. General Tweets: tweets by the public
    - Tweets at Tesla
    - Tweets at Elon Musk
    - Tweets containing Tesla hashtag (#tesla)
    - Tweets containing Elon Musk hashtag (#elonmusk)

These tweets were pulled daily directly from the Twitter API using the rtweets package available through R. By analyzing these tweets, we aim to capture people's sentiment towards Tesla. Due to Twitter API rate limiting, for general tweets, we were only able to pull tweets from October 27th onward and only a certain number of tweets were able to be accessed at a given time. Therefore, the final general tweets dataset consists of 217,193 tweets from October 27th to November 13th.

General Tweets

Date	Number of Tweets
10/27	8,894
10/28	9,147
10/29	17,988
10/30	11,475
10/31	6,064
11/01	2,650
11/02	15,323
11/03	13,393
11/04	12,028
11/05	10,856
11/06	17,979
11/07	12,470
11/08	12,232
11/09	17,995
11/10	12,044
11/11	9,820
11/12	8,546
11/13	16,969
Total	217,193

In summary, the total Twitter data available for our three categories of tweets, tweets by Tesla, Elon Musk, and the public are as follows:

Tweet Category	Date Range	Number of Tweets
Tweets by Tesla	6/7/2014 - 10/16/2020	1,189
Tweets by Elon	12/29/2019 - 11/7/2020	297
General Tweets	10/27/2020 - 11/13/2020	217,193

Noticeably, we have various date ranges and numbers of tweets available for the three categories. This makes time series sentiment analysis difficult as the date ranges do not align, and data availability caused by limitations with the Twitter API prevent further inclusion of more date ranges. In addition to the notion that tweets from Tesla, Elon, and the public may contain different content, we continued the analysis using these three datasets and analyzed the effect of each on Tesla stock market prices separately.

2.2. Data Preprocessing

In order to analyze tweets, it is necessary to clean the text data to make it more usable. This includes removing stop words, which are words that are very common but have no real meaning, such as "the," "and," "a," and "is." Next, all letters are converted to lowercase to ensure that the same letters have the same representation. For example, we want "B" and "b" to be analyzed in the same way. Furthermore, all punctuation must be removed. Lastly, words need to be stemmed, which means reducing words to the root or base word. For example, "walking" and "walked" would both be stemmed to "walk." This text preprocessing facilitates sentiment analysis and helps remove unwanted words and make text data easier to process. We used the tweet pre-processor package in python to pre process the tweets.

3. Modeling

As previously stated, because the aim of certain tweets and Twitter accounts are different in addition to issues with data availability, we split our analysis into three sections. The first section aims to predict Tesla stock prices using just tweets from Tesla's official account, @Tesla. The second model predicts Tesla stock prices based on Tweets from Elon Musk's official account, @elonmusk. The last model predicts Tesla stock prices using general tweets related to Tesla and Elon Musk but not tweeted by either official account. In summary, using these three models, we aim to predict Tesla stock prices.

The first step in using Twitter data to predict Tesla stock prices is to perform sentiment analysis. Though the date ranges differ, we performed sentiment analysis for all three subsets using the same method. After performing sentiment analysis for each dataset, we used three models to predict the closing Tesla stock price for a given day. We also lagged several variables including sentiment scores and tweet volume both lagged by up to 3 days in addition to opening, closing, and high and low prices all lagged by one day. Testing various combinations of these features, we eventually resulted in three of the best performing models for our three categories of tweets.

3.1. Sentiment Analysis

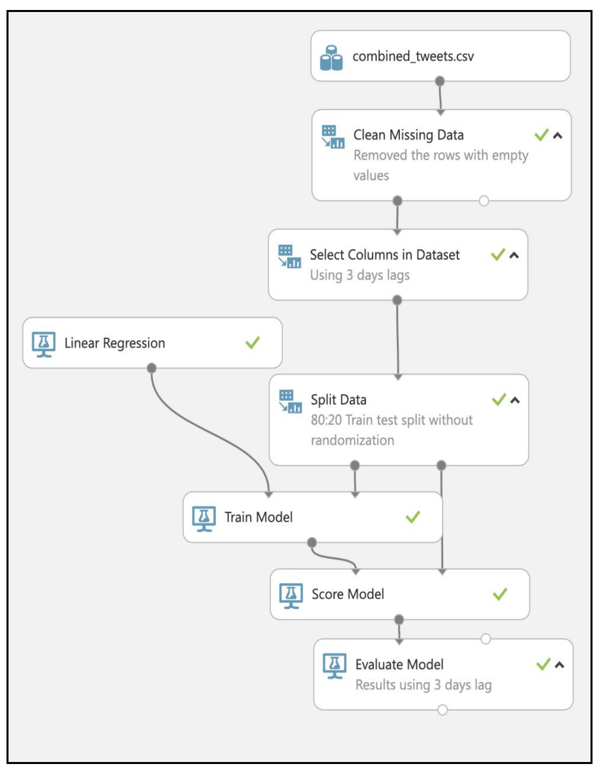
Sentiment analysis of tweets was performed using VADER sentiment analysis model, provided under NLTK package in python. VADER is a rule based sentiment analysis model which is specifically trained on social media texts, such as Twitter and Facebook. The model generates a sentiment value of [-1,1] for a given text. We generated the sentiment value of each tweet and calculated the average sentiment for a given day.

3.2. Model

After testing several models, we determined that linear regression was best fit for predicting Tesla stock price because other models did not improve prediction accuracy, and linear regression enables feature weight comparison. Furthermore, we decided to use the same linear regression model with the same features for all three datasets to facilitate result comparisons and ensure proper and consistent model comparison.

The final model was trained separately for all three datasets and follows the flowchart shown below. After cleaning missing data by deleting rows, we select the following columns:

- Sentiment lagged by 1 day: sentiment_1dlag
- Sentiment lagged by 2 days: sentiment_2dlag
- Sentiment lagged by 2 days: sentiment_3dlag
- Tweet volume lagged by 1 day: volume_1dlag
- Tweet volume lagged by 2 days: volume_2dlag
- Tweet volume lagged by 3 day: volume_3dlag
- Previous day closing price: close_prevday
- Previous day opening price: open_prevday



For each dataset, these features are used to predict the current day closing price (close). After testing 2 different lag periods of 2 days versus 3 days, model performance increased when including three-day lags, so our final model included all three prior days of data for tweet sentiment and volume. We also split our data into an 80% training set and 20% test set using a non-random split due to the temporal aspect of our data. We then trained this model using our three datasets to analyze the effect of the selected features on predicting Tesla stock price.

4. Results

4.1. Results for Model 1: Tweets from @Tesla

First, we trained the *linear regression* model described above of the dataset containing sentiment analysis for tweets from the official Tesla account, @Tesla. Our R^2 value is relatively high at 0.9978, indicating that this model is quite good at predicting Tesla stock prices based on sentiment scores from Tesla account tweets. Furthermore, both the MAE and RMSE are relatively low at 1.34 and 3.38, again confirming the model's ability to correctly predict Tesla stock prices.

Metrics

Mean Absolute Error	**1.338299**
Root Mean Squared Error	**3.384657**
Relative Absolute Error	**0.037675**
Relative Squared Error	**0.002238**
Coefficient of Determination	**0.997762**

One of the main reasons we chose to use linear regression for this project was to facilitate feature weight analysis. Analyzing the coefficient for this model, we can see that unsurprisingly, the most important coefficient in predicting the current day stock price is the previous day's closing price. Examining the sentiment weights, we can see that interestingly, the 1 day lag has a positive relationship with the closing price while the 2 and 3 day lags have a negative relationship. In addition, the 2 and 3 days lags are slightly more important, with a higher magnitude than the 1 day lagged sentiment. Furthermore, all three lagged sentiment scores are more important than lagged tweet volume, which had very little weight in predicting closing stock prices.

Feature Weights

Feature	Weight
close_prevDay	0.975901
Bias	0.244485
sentiment_2dlag	-0.0777977
sentiment_3dlag	-0.0607008
sentiment_1dlag	0.0584778
open_prevDay	0.020097
volumne_1dlag	0.0053759
volumne_3dlag	0.00080024
volumne_2dlag	0.000262582

Overall, this model that uses the tweets from the official Tesla Twitter account to predict the Tesla closing stock price performs surprisingly well. However, the actual sentiment scores do not hold a high weight in the final model.

4.2. Results for Model 2: Tweets from @elonmusk

We then trained the same linear regression model on the dataset containing sentiment scores for tweets from Elon Musk's official account, @elonmusk. The R^2 value for this model at 0.6840 was much lower than that of the Tesla model at 0.9978. Likewise, the MAE and RMSE were much higher at 8.94 and 11.44, respectively. The difference in performance between the Tesla model and the Elon Musk model may be due to data availability differences. We have much more data for the Tesla account with almost 1,200 tweets spanning 6 years. However, from Elon Musk's account, we only have 297 tweets from about the last year. This difference in data availability may make the Tesla model better at predicting stock prices.

Metrics

Mean Absolute Error	8.937803
Root Mean Squared Error	11.435066
Relative Absolute Error	0.576079
Relative Squared Error	0.316023
Coefficient of Determination	0.683977

Interestingly, analyzing the coefficients for this model showed some surprising results. First, the most important feature in predicting Tesla closing price by a significant amount was the 3 day lagged sentiment scores from Elon's account. This feature, sentiment_3dlag, had a feature weight of 6.28, which represents a relatively high, positive relationship with the sentiment score and the Tesla closing stock price. However, the second most important feature is the 2 day lagged sentiment score, sentiment_2dlag, which conversely has a negative relationship with the closing price. Lastly, the 1 day lagged sentiment score had a small weight. As in the Tesla model, the previous day's closing price fell somewhat high up in terms of coefficients in this model as well, with a 1.17 weight. Likewise, Elon's tweet volume does not seem to be important in predicting Tesla closing stock price.

Feature Weights

Feature	Weight
sentiment_3dlag	6.28167
sentiment_2dlag	-1.37191
close_prevDay	1.17173
sentiment_1dlag	0.416582
open_prevDay	-0.17228
Bias	-0.140404
volumne_1dlag	0.0815415
volumne_2dlag	-0.0639025
volumne_3dlag	0.0527602

Overall, this model that uses Elon Musk's tweets to predict the Tesla closing stock price performs relatively well but not as well as the prior model using Tesla tweets.

4.3. Results for Model 3: Tweets About Tesla and Elon Musk

Lastly, we attempted to fit the same model to our dataset containing over 217,000 from the general public, including tweets at Tesla and Elon and tweets with the Tesla and Elon Musk hashtags. However, model results were extremely inconsistent, resulting in very negative R^2 values. This indicates that with the data available, the model was not at all able to predict the closing price.

This poor model performance on this dataset can be due to a variety of reasons. First, though there were many more tweets available in this dataset, the date range was much smaller. Accounting for weekends and after lagging variables, we essentially only had 10 meaningful days of data even though tweets were pulled daily for almost three weeks. Second, there is much more variability in these tweets than tweets stemming from just Tesla or Elon Musk's official accounts. People may "at(@)" or hashtag Tesla and Elon Musk for a variety of reasons, not just to talk about Tesla itself. This can lead to a significant amount of variability in the sentiment scores for this dataset.

Based on our prior analysis of Tesla and Elon Musk's official accounts, it is likely that we would have needed at least several months of data to achieve meaningful results.

4.4. All tweets combined

Though we initially planned on just using general public tweets to predict Tesla stock prices, our analyses on Tesla and Elon Musk's tweets showed promising results. Thus, we decided to combine these three datasets, creating a dataset with Tesla and Elon Musk's sentiment scores in addition to the sentiment scores from the tweets by general users. We ran the same linear regression model to predict the closing price. We see that the R^2 value of 0.9538 falls between the prior two R^2 values but is much closer to the Tesla model's R^2 value. However, the MAE and RMSE are somewhat high at 11.43

Metrics

Mean Absolute Error	11.42725
Root Mean Squared Error	27.500347
Relative Absolute Error	0.096844
Relative Squared Error	0.046241
Coefficient of Determination	0.953759

and 27.50. Overall, this model that combines Tesla and Elon Musk's official account tweets performs relatively well.

Analyzing the coefficients for this model, we again see that the previous day's closing price is the most important feature in predicting the current day's closing price. Furthermore, the 3 day lagged sentiment score has more weight than the 2 and 1 day lags. Interestingly, the 3 day lagged sentiment score has a negative relationship whereas the 1 and 2 day lagged sentiment scores have a positive relationship. Again, the tweet volume seems to be relatively unimportant, as it consistently has a low weight in all models.

Feature Weights

Feature	Weight
close_prevDay	1.04956
Bias	0.313736
sentiment_3dlag	-0.161989
open_prevDay	-0.0546986
sentiment_2dlag	0.0227533
sentiment_1dlag	0.0189411
volumne_3dlag	-0.0157668
volumne_2dlag	0.0041298
volumne_1dlag	0.00390814

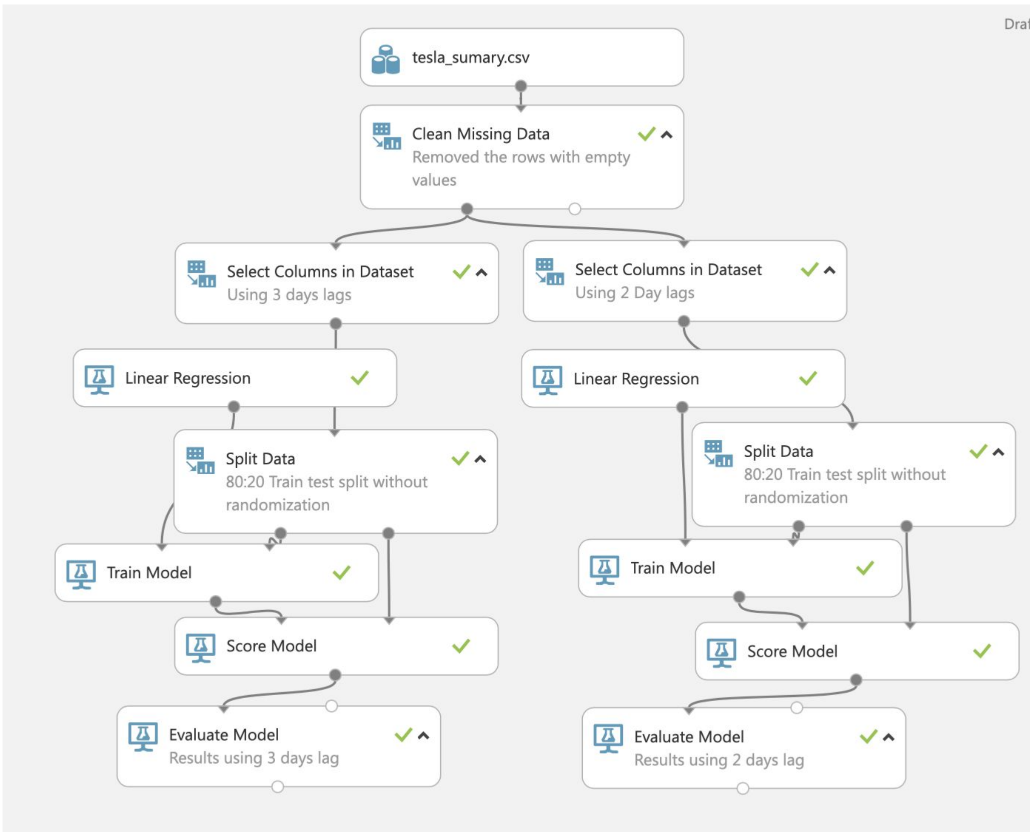
Overall, this model that combines tweets from Tesla and Elon Musk with tweets from the general public about Tesla and Elon Musk performs relatively well.

5. Conclusion & Remarks

Based on the results from our three models, we cannot necessarily say that using sentiment scores was extremely beneficial in predicting Tesla stock prices. Though all models performed relatively well, data availability issues and differences in date ranges make it difficult to directly compare results. Analyzing the coefficient weights, the variability in whether sentiment scores for different lagged periods had a positive or negative weighting with the Tesla closing stock price indicates that sentiment scores may be inconsistent. For example, in the last model that combined datasets, the 3 day lagged sentiment score had a negative weight but in the 1 and 2 day lags had a positive weight. This potentially shows that sentiment scores vary significantly by day and lag period, and thus affect the model's ability to predict closing prices differently. Furthermore, all models found the previous day's closing price to be one of the most important predictors. While this is not surprising, it indicates that using sentiment scores may not enrich the model by a significant amount depending on the dataset of interest. The main case where sentiment scores did seem to have an impact was when analyzing Elon Musk's tweets. This is interesting because Musk tweets about a wide variety of subjects, not just about Tesla.

Overall, this project was very interesting in starting a discussion on using social media to predict stock prices. While Elon Musk is a particularly vocal and active CEO on Twitter, it seems that more research needs to be done to better quantify the effect of tweets on stock prices. Future analyses could analyze tweets for a longer period of time, which would most likely improve model performance, as it was one of the biggest challenges in this project. Furthermore, using natural language processing to analyze tweet content could also be helpful in better predicting Tesla stock prices. All in all, the use of social media platforms like Twitter to predict stock prices are an interesting topic worthy of further research.

6. Appendix



References

1. https://www.wsj.com/articles/tesla-stock-price-still-makes-no-sense-11603360922
2. https://www.cnbc.com/2020/09/06/tesla-could-be-the-most-dangerous-stock-on-wall-street-investment-researcher-says.html
3. Elon Musk tweet sends Tesla's stock plunging - The ...www.washingtonpost.com › musk-tesla-stock

GitHub

https://github.com/akshaypunwatkar/tesla_stock_prediction