Into the Twitterverse

An Analysis of Social Media Sentiment as an Indicator for Stock Price

Adwith Mukherjee, Maggie Beardsley, Serge Belin, Deniz Ozturk

Background

With commission-free investment in the stock market on the rise, social media users of all backgrounds are posting their feelings about their investments. We chose to explore the relationship between one of the most ubiquitous investments of 2020, TSLA, and the sentiment of the general public concerning the company on Twitter and Reddit

Hypothesis

On a given day, the sentiment of social media mentions of Tesla stock on Twitter and Reddit can be used as an indicator of Tesla stock price performance.

Data & Collection

Twitter:

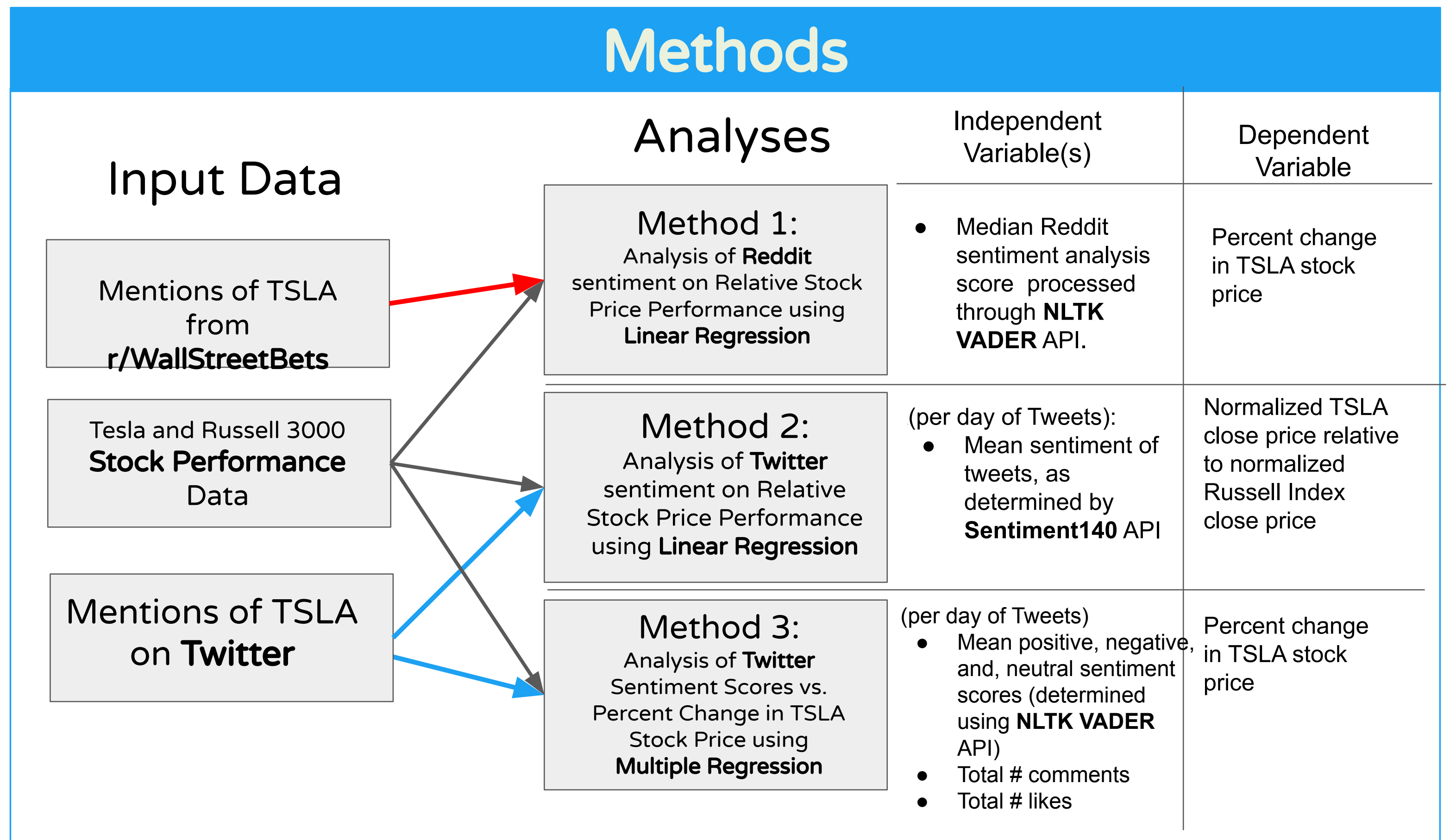
- A database of 13,000 TSLA-related Tweets spanning from January 1, 2020 to October 3, 2020 was scraped directly from Twitter using Selenium. Tweet **body, comments, and likes** were saved.
- The input for the multiple regression analysis was constructed as follows:
  - Each Tweet was cleaned to remove embedded hyperlinks and user mentions. Then, each Tweet was passed into a NLTK **VADER Sentiment Analyzer** to yield a vector of the **positive, negative, and neutral sentiment** scores of each tweet.
  - This database was reduced by date to create a dataset of the mean positive, negative, and neutral **sentiment scores for tweets on a given date**, as well as the total number of comments and likes.

- The input to the linear regression analysis was constructed as follows:
  - Each tweet is passed into the Sentiment140 API and a sentiment score is assigned on a scale of 0 (most negative) to 4 (most positive).
  - For each day, the sentiment scores of all the tweets on that day are averaged.
  - The relative performance of TSLA stock compared to the Russell index fund 30 days after the Twitter posts is computed.

Reddit Posts:

- We used the pushshift Reddit API to retrieve posts from wallstreetbets that mentioned $TSLA from January 1st 2019 to December 31st 2020.
- This database was used in combination with investing.com json data. We processed the sentiment using NLTK with some Reddit specific vocabulary.

Methods



Input Data

Mentions of TSLA from **r/WallStreetBets**

Tesla and Russell 3000 **Stock Performance** Data

Mentions of TSLA on **Twitter**

Analyses	Independent Variable(s)	Dependent Variable
Method 1: Analysis of **Reddit** sentiment on Relative Stock Price Performance using **Linear Regression**	• Median Reddit sentiment analysis score processed through **NLTK VADER** API.	Percent change in TSLA stock price
Method 2: Analysis of **Twitter** sentiment on Relative Stock Price Performance using **Linear Regression**	(per day of Tweets): • Mean sentiment of tweets, as determined by **Sentiment140** API	Normalized TSLA close price relative to normalized Russell Index close price
Method 3: Analysis of **Twitter** Sentiment Scores vs. Percent Change in TSLA Stock Price using **Multiple Regression**	(per day of Tweets) • Mean positive, negative, and, neutral sentiment scores (determined using **NLTK VADER** API) • Total # comments • Total # likes	Percent change in TSLA stock price

Twitter Analysis

Multiple Regression with VADER (Method 3)

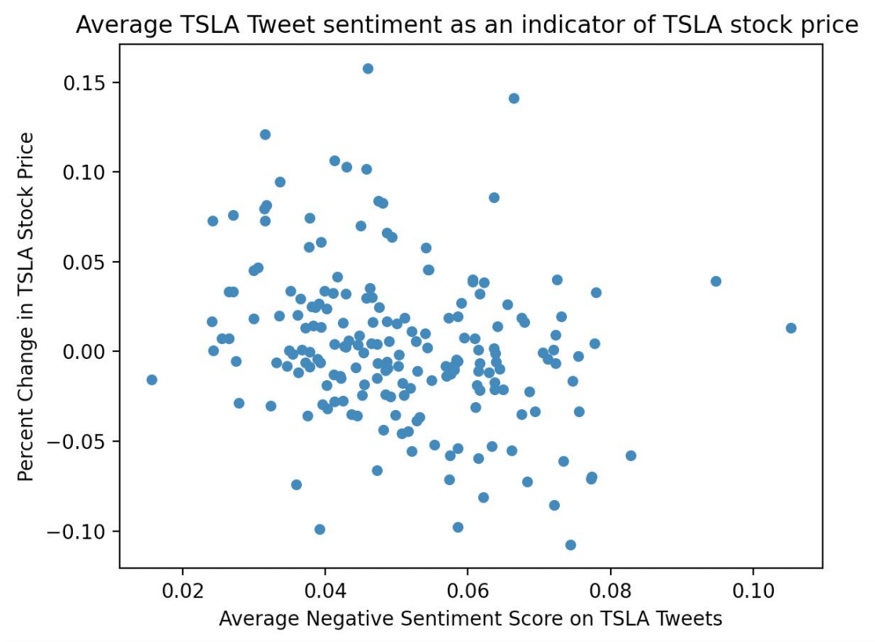


Figure 1: Relationship between the most significant independent variable: "neg" and "pc".

We ran an ordinary least squares multiple regression test to find the relationship between the average positive, negative, and neutral sentiments ["pos", "neg", "neu"] as well as the mean # of comments and likes ["comments", "likes"] on Tweets published each day with the change in stock price on that day ["pc"]. First, to isolate the most significant variables, we determined their correlation constant with "pc" (Figure 2) and used these in the regression analysis (Figure 3).

Independent Variable	Pearson's r
"pos"	0.190
"neg"	-0.270
"neu"	0.0549
"likes"	0.055
"comments"	0.061

Figure 2: Pearson correlation constant of each proposed independent variable. Those in green ("pos, "neg") were selected for regression.

Ind. Var.	Coeff	P-value
"pos"	0.8879	0.002
"neg"	-1.0335	0.001

p-value	1.29e-5
R^2	0.213

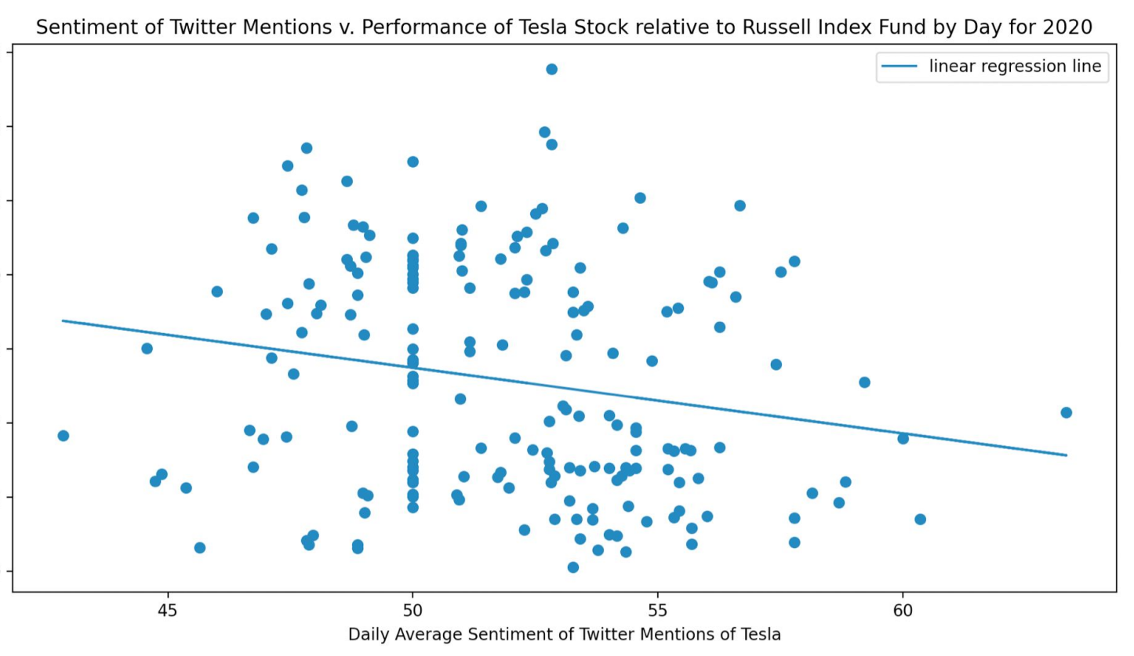
Figure 3: Top: "pos" is positively correlated with "pc", while "neg" is negatively correlated. Both have significant p-values, along with the overall model.

We discovered significant results for two variables: "pos" and "neg". There is no evidence to suggest that dates with greater Twitter activity (dates with higher mean likes and comments) influence stock price on that date.

We can conclude that **there is a significant relationship** between the average sentiment values for a day of Tweets related to Tesla and the share price of the company.

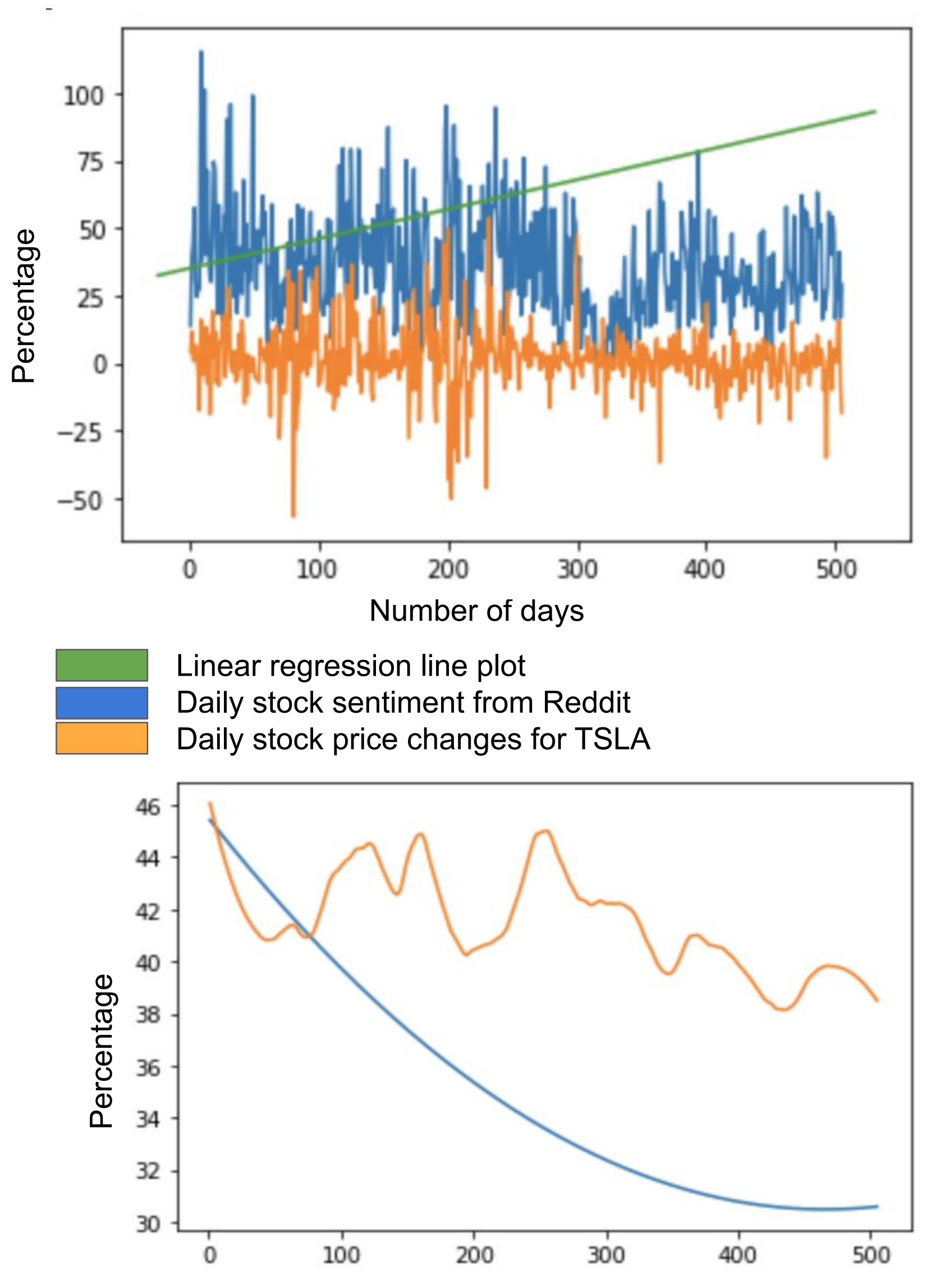
Simple Linear Regression with Sentiment140 (Method 2)

The results of the linear regression with sentiment computed based on Tesla Twitter mentions resulted in a p-value of about 0.008 and an r-squared of 0.04. The p-value is significant so the null hypothesis can be accepted and the alternative hypothesis, that Twitter sentiment can be used as an indicator of stock performance 30 days later, is accepted. However, it is important to note that the r-squared value shows that the linear regression model is not necessarily a good fit for the data.To the left is a graph showing sentiment on the x-axis (the independent variable of the regression) and relative stock performance on the y-axis (the dependent variable of the regression). The dots show the data points used for the regression model, and the blue line shows the linear regression line calculated by the statistical test. There is a clustering of values with an average sentiment of 50 because the API used classifies many posts as neutral.



Reddit Analysis

Graphs representing linear regression between daily sentiment from Reddit and daily stock price movement for TSLA



For unfiltered data:

p-value	0.12
R squared	0.07

The r-squared and p value found from linear regression showed us that the two datasets from stock sentiments and daily price movement were not dependent on each other in a statistically significant way.

However, since the linear regression plot has a slope slightly bigger than 0, there is some correlation between the two datasets but the data is underfitting based on the standard error.

Social media as a financial platform is a relatively new phenomenon, and as more data can be collected over time under different market conditions, this experiment can be successful.

After running our data through a salgov filter to remove noise, we saw that the actual changes in stock price were a lot more stagnant than sentiments. Even though both the stock price movements and sentiments are hinting a stagnation period or a drop in price there is no significant correlation to prove it.

In the graph on the left the stock price change is shifted up to start from the same horizontal point through time, and we can see a significant drop in sentiment with continuous, slower drops in price changes, possibly hinting a drop in volume.

Conclusions

The null hypothesis can be rejected and the alternative hypothesis can be accepted: the sentiment of social media mentions of Tesla on Twitter can be used as a same-day indicator of Tesla stock price performance. Although we were not able to reject the null hypothesis when tested on Reddit data, future exploration into denoisifying data could reduce the p-value into the range of significance.

Challenges

- Restructuring data to meet our needs, specifically retrieving posts from social media sites, aligning this data with stock price data, and cleaning data for statistical testing
- Reducing noise in social media data
- Our analysis had dependencies on NLP API's, which have known shortcomings in accuracy