Predicting Stock Market Prices using Sentiment Analysis and Machine Learning Algorithms

John Fiester, Faculty Advisor: Chantal D. Larose
Eastern Connecticut State University

Abstract

This project represents the first phase of predicting stock market prices using sentiment analysis of tweets and financial news and machine learning algorithms.

The project is motivated by the concept that stock prices are driven by investors' attitudes and on new information.

Sentiment analysis allows us to quantify the tone or emotion behind certain pieces of text and media.

The **goal** is that with sentiment data collected through tweets and financial news, along with stock prices of several businesses, we can build a multiple linear regression model in attempt to describe the data, its potential trends, make price predictions and to give an idea of how other models may preform.

Data Collection

Stock price data was gathered for several stocks and one stock index from Yahoo Finance [2]. Figures in this presentation indicate findings for Tesla's stock (TSLA). It is important to note that results and data varies between different stocks.

Tweets were collected through the Twitter API [3] by using a search query for each stock. Tweets were collected each day and do not include retweets.

Financial news was collected through Reuters [4] and through Market Watch [5]. News articles were collected each day.

Exploratory Data Analysis

Tweet and news data were cleaned (removal of links, punctuation, stop words, etc.) allowing smooth sentiment analysis.

From sentiment analysis, ten emotion variables were created along with variables indicating mean score per each emotion per days worth of tweets and news articles.

Figure 1 depicts the mean scores for each emotion for Tesla tweets and shows how some emotions tend to move together over time.

Based on Figure 1 it appears positive emotions (Joy, Trust, etc.) may have high correlation with one another, whereas the same applies for negative emotions (Disgust, Sadness, etc.)

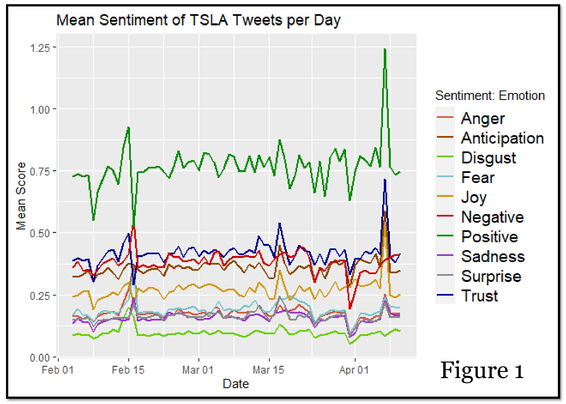

Figure 1

Principal Component Analysis

As expected from intuition and Figure 1, some emotions tend to be highly correlated with one another, causing multicollinearity

Multicollinearity reduces the precision of the estimate coefficients, which can weaken the statistical power of the model.. Multicollinearity was addressed using **Principal Component Analysis** (PCA).

PCA reduced the initial number of variables from 20 highly correlated variables to four independent components through linear combinations that contain most of the information from the initial variables.

For the Tesla data, the grouping of variables through PCA resulted in four components namely; News, Negative Tweets, Positive Tweets, and Positive News. Figure 2 depicts the breakdown of variables included in each component.

From Figure 2:

Component 1 (RC1) is the News Component.

Component 2 (RC2) is the Negative Tweets Component.

Component 3 (RC3) is the Positive Tweets Component.

Component 4 (RC4) is the Positive News Component.

```
Loadings:
       RC1    RC2    RC3    RC4
AngMN  0.862
AntMN  0.609                  0.667
DisMN  0.780
FearMN 0.891
JoyMN                         0.854
SadMN  0.879
SurMN                         0.696
TruMN  0.594                  0.717
NegMN  0.908
PosMN  0.645                  0.707
AngMT          0.915
AntMT                 0.893
DisMT          0.905
FearMT         0.950
JoyMT                 0.965
SadMT          0.900
SurMT          0.810
TruMT                 0.954
NegMT          0.961
PosMT                 0.957
```

Figure 2

Regression Model

A **multiple linear regression** (MLR) model was constructed to predict the next day's stock prices using the PCA component scores.

Descriptive regression model as we are looking to the model to see potentially how other models may perform but not to make inferences about a population.

Model explains about 26% of variance and fitted values in Figure 3 show directional promise.

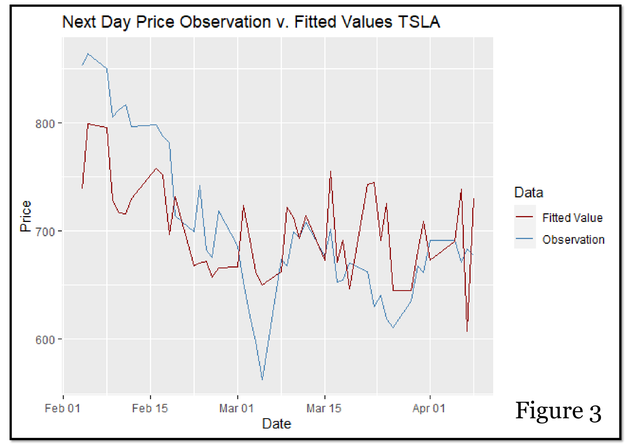

Figure 3

Conclusion & Further Research

The modeling process provokes the idea of comparing a MLR model with a smaller sample window that shifts for each new observation rather than aggregating all data. Data that is less recent may punish the model more than it helps improve it as stock prices can change drastically through larger periods of time.

Future research includes constructing models by using other machine learning algorithms and evaluating their performance. Model shows potential promise for more powerful algorithms.

Algorithms such as logistic regression, artificial neural networks, naïve bayes, and support vector machines will be used to create a set of models so that they can be compared.

References

[1] Larose, D. T., & Larose, C. D. (2015). *Data Mining and Predictive Analytics (Wiley Series on Methods and Applications in Data Mining)* (2nd ed.). Wiley.
[2] *Telsa, Inc. (TSLA).* (n.d.). Yahoo Finance. Retrieved April 12, 2021, from https://finance.yahoo.com/quote/TSLA/history?p=TSLA
[3] *Twitter API Documentation.* (n.d.). Twitter Developer. Retrieved April 12, 2021, from https://developer.twitter.com/en/docs/twitter-api
[4] *Reuters Editorial.* (n.d.). Business & Financial News, U.S & International Breaking News. Reuters. Retrieved April 12, 2021, from https://www.reuters.com/
[5] *Latest News.* (2021, April 14). MarketWatch. https://www.marketwatch.com/