The Influence of Twitter on the Stock Market

Rolando Franqui and Connor Brown

5/1/2020

Main Research Goal:

> Can we Predict the Stock Market Performance using Twitter Data?

Introduction

In recent years, social media has become a big part of our daily lives; it affects how we communicate with others, our jobs and the economy. With social media platforms such as YouTube, Facebook, and Twitter becoming extremely popular, people are allowed to express and share information that might not be necessarily accurate or merely just an opinion; however, having such a wide reach can affect many areas significantly. The goal of our research project is to see to what extent social media (specifically twitter) has affected the behavior of traders and the economy, despite being in the midst of an economic crisis. To analyze the tweets we used sentimental analysis, we will calculate the mean sentiment of the day and compare it to the adjusted close price of the S&P 500 Index we used as our metric for market performance.

Data collection

We first started by manually searching twitter using the built-in search engine by passing in keywords that we might be able to use to gather market data. This was an extremely useful part of the process because it gave us an idea of what kind of data the queries could bring us. This process helped us eliminate queries that would add substantial noise to our data due to ads or unrelated tweets that that query gathered. We eventually settled on 8 different queries. With the use that we gathered using the following queries: “S&P500”, “Economy”, “SPY”,“VOO”, “Stimulus Check”, “Crude Oil”, “Stock Market”, “Recession.” From this sentiment analysis, we will calculate the mean sentiment of the day and compare it to the adjusted close price of the S&P 500 Index we used as our metric for market performance.

Getting a Sense of the Data

To get a better understanding of the data, we created a word cloud of the most popular words from the tweets we collected from the past four months. The words that are shown here have a frequency of at least a hundred inside our collected tweets. Some of the terms that caught our eye were ‘Trump’, ‘Obama’, ‘Coronavirus’, ‘Covid’, ‘Unemployment’, ‘China’, ‘Virus’, ‘Fall’, and ‘News’. While none of these were keywords used to scrape the tweets, they do represent the current climate in society, and show what current events might have an effect on the market.

Performing sentiment analysis

Sentiment analysis is a type of data mining that measures the inclination of people's opinions through natural language processing (NLP), computational linguistics and text analysis, which are used to extract and analyze subjective information from Social Media or other mediums such as text documents. In this case we are using it to analyze the sentiments of the tweets we collected. In the figure below we created a bar plot of the sentiment of the tweets broken down in columns these include one for each emotion type as well as a positive or negative valence. The ten columns are as follows: "anger", "anticipation", "disgust", "fear", "joy", "sadness", "surprise", "trust", "negative", "positive."

The most prevalent emotion seemed to be trust, and the most common sentiment seemed to be negative. The emotions seemed to be reasonably balanced, most probably due to the market being at an almost all-time high at the beginning of the year and then going on a very steep decline towards the second half of our data. This leads to to belive there is a correlation between the market and our twitter data.

Market Performance vs Sentiment Analysis

To better observe if market performance was in line with the sentiment corresponding to the same date, we calculated the mean sentiment of every day of tweets using sentiment analysis. We plotted the results in a time series and displayed it next to the stock's performance and making the dates line up (Figure Below). What we observed when we did this is that, in fact, the sentiments correlated with the market performance. We also saw that twitter is reacting to current events. We can extrapolate this because the significant changes in sentiment seem to happen a day after a significant difference in the stock market performance.

Creating a linear model

A linear regression model is a model that assumes a linear relationship between the input variables (x) and the single output variable (y) Example: $Y=B0+B1*xY=B0+B1*x$.
Our linear model is summed up in the following equation $ClosingPrice=a(Intercept)+AvgSentimentofDay\times b1$ClosingPrice=a(Intercept)+AvgSentimentofDay×b1. So, if we were to assume that in a random day the Average sentiment is -.30, we would expect the closing price to be on average around 2746.538. The equation would look the following way $3070.67+(1080.44+-0.30)=2746.53$3070.67+(1080.44+−0.30)=2746.53. If we look at our R-Squared Value of 0.20, this means our model can account for 20 percent of the variance exhibited in the data. This is not exceptional but acceptable in our view due to the fact that we only have a single predictor variable in an extremely unpredictable market segment.

Attempting to Predict the Market with Twitter Data

After we assessed the correlation between the S&P 500 Index and Twitter sentiment, we used the data to try to predict the index price. To make these predictions, we utilized our linear model and the sentiments exhibited during the tweets in April to obtain the predicted index closing price at each corresponding market day. We created a testing data set, which includes a subset of the data containing the mean sentiments and adjusted closing prices from the month of March, and we ran it against a testing set ,which was a subset of our data containing the sentiments of April.

After analyzing the model, we came to the following conclusions. From our obtained R-Squared Value, our model can only account for 5 percent of the variance and has a P-Value of 0.152, meaning our

findings are not very accurate. This can be attributed to the extremely volatile market we have seen during the past two months. We observe the market shapely rising in April while it was crashing in March. This and the fact that we are only using one variable as a predictor could explain the lack of accuracy in our model.

Conclusion

In closing, we discovered that there is, in fact, a moderate correlation between daily tweet sentiment and stock market closing price. Through a step-by-step plan of 1.) Web scraping tweets related to the market using a Python script, 2.) using word processing methods to find words that exhibit sentiment, 3.) using these words to calculate a value of positive or negative sentiment towards the market, and 4.) comparing each day's sentiment to its corresponding day's adjusted closing price in S&P 500, we found a correlation between daily tweet sentiment and stock market adjusted closing price differences. After the fact, we attempted to use these discoveries to predict S&P 500 Index closing prices based on daily twitter sentiment. Using the month of March as a training set, we created a linear model and used it to predict adjusted closing prices for April, with inconclusive results. When we compared the predicted April prices to the actual data we had, we found a p-value of just below .05, which, while technically significant, led us to believe that we do not have enough parameters to predict prices accurately. The predicted prices are not exceptionally poor but are not enough to be called an accurate prediction method. We believe that the use of more sentimental analysis from other websites, such as news headings, Facebook, and other sources with the addition of multiple linear regression, could further propel these predictions closer to the actual data.

Sources and Libraries

We utilized the following Libraries:

- (tm) #Cleaning the Tweets
- (wordcloud)#Creating the wordCloud
- (syuzhet)#Sentiment Analysis
- (sentimentr)
- (lubridate)#Sentiment Analysis
- (ggplot2) #Plotting
- (scales)#Plotting
- (reshape2)#Plotting
- (dplyr)#Plotting
- (gridExtra)#Joining graphs together
- (SnowballC)#Creating the wordCloud
- (RColorBrewer)#Creating the wordCloud
- GetOldTweets3#Python data Scrapping Library

Other resources:

1. https://www.datacamp.com/community/tutorials/linear-regression-R

2. https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html
3. https://medium.com/swlh/exploring-sentiment-analysis-a6b53b026131
4. https://towardsdatascience.com/how-to-scrape-tweets-from-twitter-59287e20f0f1
5. https://finance.yahoo.com/quote/%5EGSPC/history/
6. https://rpubs.com/thirus83/453609
7. https://www.r-graph-gallery.com/ggplot2-package.html
8. https://github.com/Jefferson-Henrique/GetOldTweets-python/tree/master/got3