Twitter Sentiment and the Market

Talon Marquard
Honors Program Senior Project
Department of Computer Science
University of Wyoming
Email: tmarquar@uwyo.edu

Abstract

The stock market is complex, and influenced by a mixture of machines, amateur traders, professional traders, executives, and the government. By analyzing the market with a variety of methods, it is hoped to find a model that can give some information regarding its trends. First by using an ARIMA model to search for any overarching trends, and then by using Twitter to learn the sentiment of people regarding the market, it is hoped to see what influence sentiment has upon the percent change in the market on a given day. As expected, the market, people, and human language processing are quite complex, especially for a computer. Much more work and refinement will be required to discover true trends.

I. Introduction

Today's stock market has been changed by technology as much as everything else, if not more so. Modern computing has also allowed for advanced algorithms to make many trades based upon micro-trends. This has become a profitable line of business and has contributed to the great volatility in the market, since the buying power of these algorithms is so great. Another increase in the volatility is that almost anyone can now put their money into the market. The Internet has allowed for almost instant trading from all over the world. Though an individual's buying power may be small, so many people are now invested and able to make fast trades, that they have a significant impact on market trends.

The web is an increasingly busy and trafficked place. The prevalence of social media posting allows for analysis sentiment. Opinion mining is a quickly growing area of data science. A convenient place for gathering data is Twitter, since posts are generally textual, and limited to a certain number of characters. Posts are also frequently tagged by the person posting. These tags allow an easy way to search through posts. Twitter also has an API that allows for easy retrieval of tweet data. All of this means that it is possible to sample tweets and possibly get an accurate idea of the sentiment around certain stocks and/or the market in general. This sentiment is likely to have some monetary impact on the market. A model built around this idea would be a powerful tool for understanding our modern digitally influenced economy.

II. Related Research

Many statistical market models have been developed over the years. Statistical algorithms such as artificial neural networks, support vector machines, and random forests have been used to try and predict stock price [4]. ARIMA models have also been used to try and detect market trends and to try and predict where the market is heading [6]. These are old techniques that are constantly being refined and improved with additional information and better data sampling. They work with certain degrees of

success, some with accuracy up to 77% [5]. ARIMA modelling is particularly useful for analyzing micro trends and micro-trading. ARIMA models allow for a strong degree of accuracy in the short term, but quickly lose predictive power over time. Some of these techniques have been put together to again improve the reliability of the models [6]. The biggest thing all these models look for are underlying trends and patterns inherent in market data. All this implies a belief that the market is predictable in of itself, but as we know the market is influenced by much more than just itself. There are economic and political changes that can cause the above models to fall apart quickly since the market has these outside influences. ARIMA models can take these critical events into account once they have happened, but, as far as I know, not as they are happening.

Twitter is a useful resource for sampling opinions. It provides large amounts of tagged data and has a large variety of users that represent all demographics of the United States and more [3]. One can then train a model to read sentiment and the subjectivity of the posts. Doing this requires feature extraction. URLs, usernames, and emoticons that are not valuable and can be removed to make it easier to get key words [3]. Sentiment analysis is also working to branch out into multiple languages so that a greater number of tweets can be sampled [3].

III. Data

To get the data for my market model, I used the Quandl API to get Nasdaq data. Quandl allows for easy retrieval of large amounts of day to day market data. It provides day high, low, index value, total market value, and dividend market value. For the ARIMA and regression model only the index values from December 2, 2017 to October 19, 2018 are used. The index data was transformed so that the day's percent change is calculated, since I am interested in market movements rather than the value itself. This calculation causes a data point (December 2, 2018) to be lost, so this end ups being 221 days of data (market is closed weekends and holidays).

The same date range of twitter data is gathered from the news source CNBCnow's twitter feed. CNBCnow was selected since it provides quick headlines of daily events that may be relevant to the market. This data was then run through Textblob, a sentiment analysis API. This gives me sentiment and subjectivity for each tweet. However, CNBCnow tweets multiple times a day, every day. To be able to examine a day to day relationship between tweet sentiment and subjectivity I transformed the data so that there is an average sentiment and subjectivity each day. I also kept the number of tweets each day to see if that may be an additionally beneficial feature.

IV. ARIMA Modelling

An ARIMA model is a time series and forecasting tool. For this I used the raw index data to begin with. I began by examining the plotted data as well as the ACF and PACF.

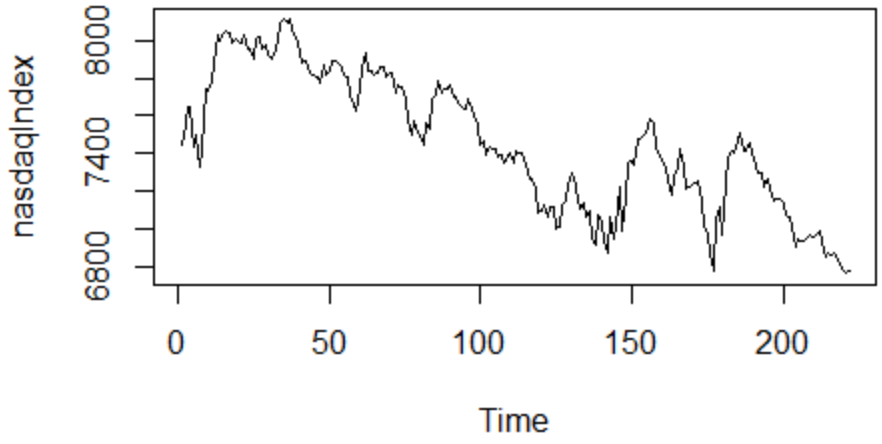

Figure 4.1: The time series plot of the Index from December 2, 2017 to October 12, 2018

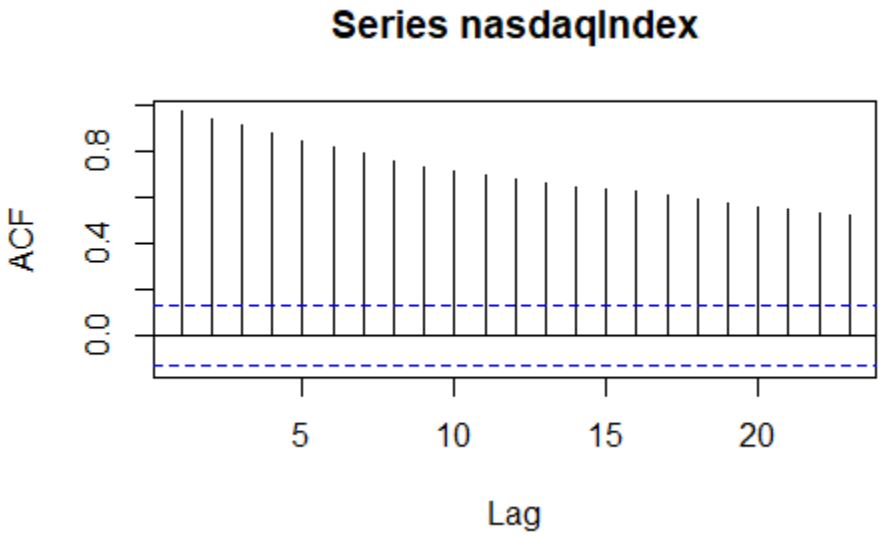

Figure 4.2: The ACF plot of the Nasdaq index values

Since the lines seem to have a fairly linear trend a difference of the data will allow us to see real patterns.

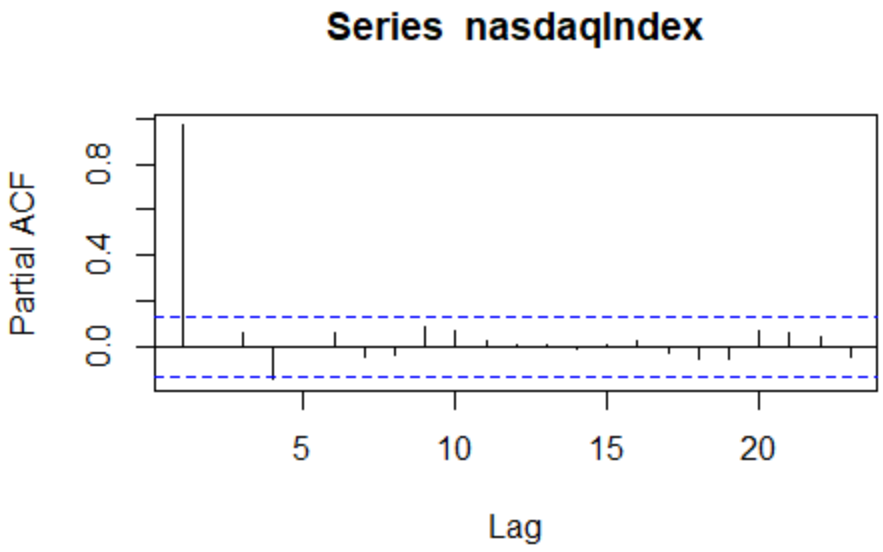

Figure 4.3: The PACF of the Nasdaq index data

From Figure 4.1 and 4.2 it is clear that there is a linear trend in the data, so a difference transformation will be used to then do further analysis. Figure 4.1 also looks as if it may be improved with a log transformation, which makes sense because stock data is usually evaluated by percent change, rather than straight dollar value change.

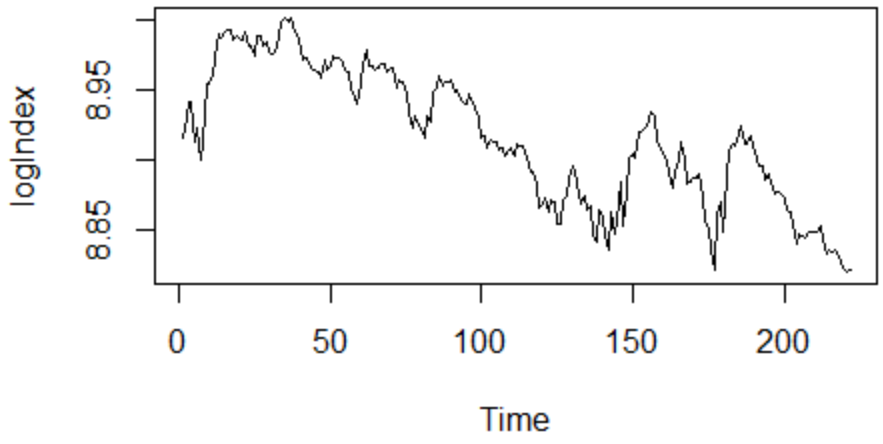

Figure 4.4: Log transformation of the time series plot of index values

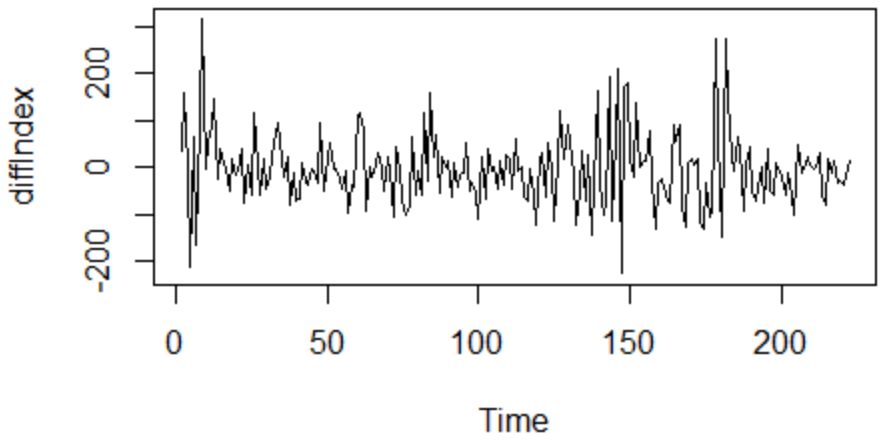

Figure 4.5: Log transformation and difference of the time series plot of index values

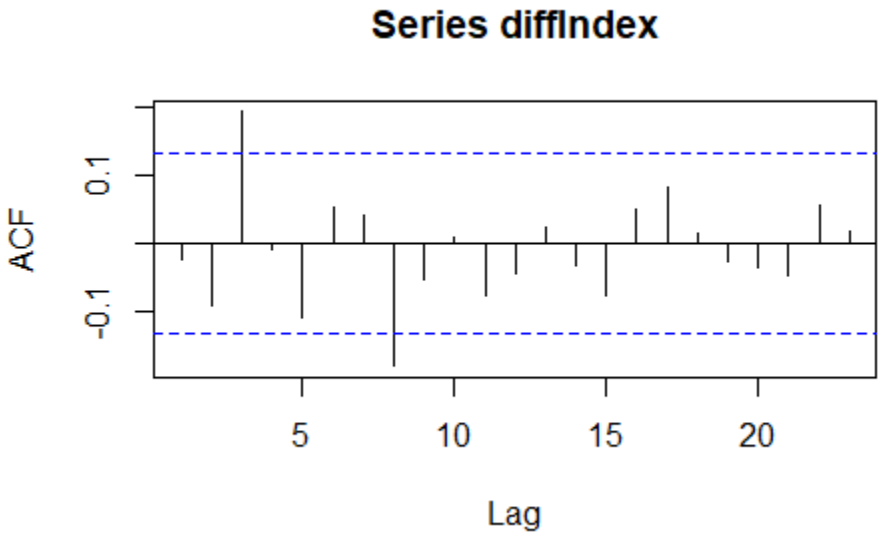

Figure 4.6: The ACF plot of the logged and differenced Nasdaq index values

The points of interest here are at lag 3 and 8. They are slightly past our threshold line. Unfortunately, there is not much that can be interpreted from this since the deviation is slight and not at any significant points.

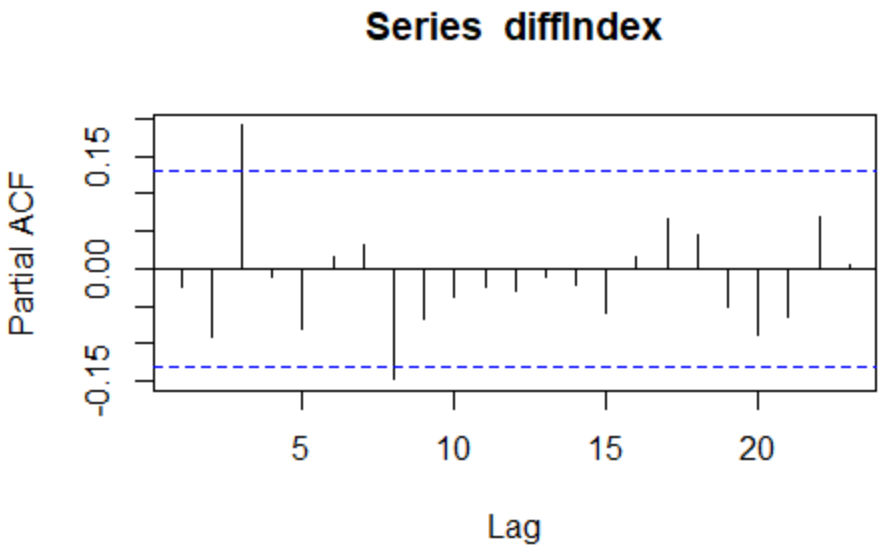

Figure 4.7: The PACF plot of the logged and differenced Nasdaq index values

The PACF is almost identical to the ACF. Unfortunately, it also gives almost no useful information.

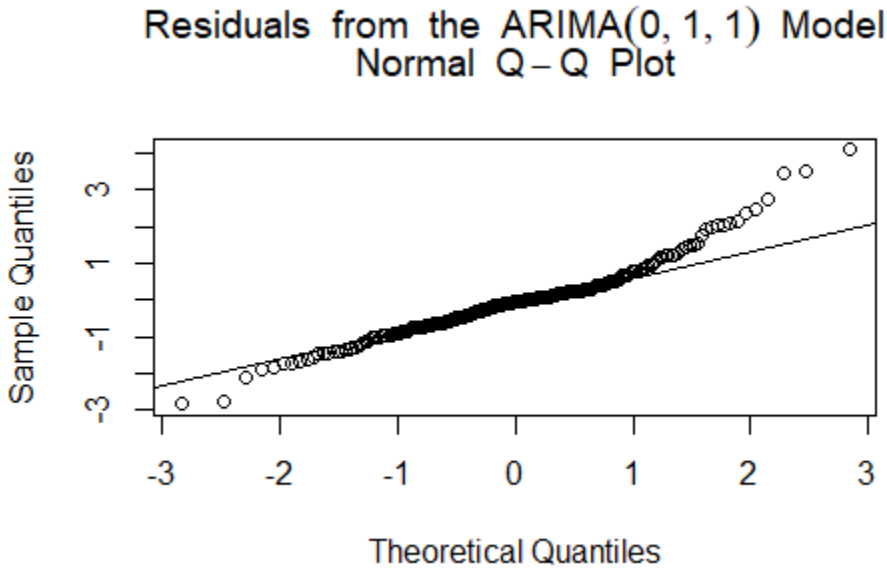

Figure 4.8: Normality plot

The transformed data leaves me with very little information. Since the PACF (figure 4.7) and ACF (figure 4.6) only have 2 points each that only barely pass the threshold line of interesting, it is difficult to determine any underlying patterns in the data. The model with the best information criteria is the simple I(1) model, which means that the movements of the data trend in a direction, but have no real underlying pattern. The IMA(1,1) model and the ARI(1,1) model both have worse information criteria

and provide little to no valuable information. To make things even worse for this analysis the assumption of normality appears to be a very poor one since points quite obviously trend way from the normal line in figure 4.8. The results of the ARIMA test models reveal no helpful information.

It is possible that the data is too numerous and too spread out. Perhaps some trends would reveal themselves more on the short term, or in higher sample rates, like trends of a single day from minute to minute. However, given the data I am working with it becomes clear that ARIMA is going to be of little to no help.

V. Regression Analysis

From the ARIMA analysis we get very poor results. It is also worth noting that sentiment played no part in that analysis. ARIMA modelling looks for trends in the data and it is difficult to include outside information in the model. Since ARIMA led me already to consider differenced and logged data, it seems that perhaps I could use that data to see if sentiment plays a part in the days percent change. Using a regression model seems like a good choice for testing for correlations in data. The reason it is important to phrase data correctly is that regression can only inform one about what is likely to be true given data from the past. It has no ability to predict into the future or outside the scope of previous experience. So, by looking at correlations between percent change and days sentiment, I may be able to determine if a certain sentiment is correlated with a certain change in the market. This means I cannot say what will happen tomorrow, but rather that if sentiment is X, then the market change is likely to be Y.

From the below figures, we see that there are no clearly evident trends between any of the variables. They all seem randomly clustered with no perceptible trend of pattern. This is not promising for our model, but it is possible that there is something slight that I am not noticing or may be revealed in combination. Though are model isn't huge, it is sizable enough to have reasonable power to detect differences and trends.

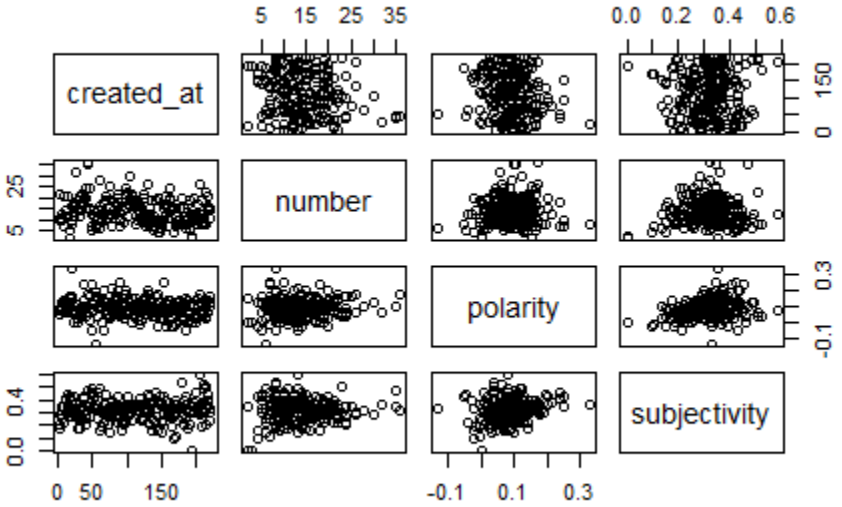

Figure 5.1: Predictor variables plotted against each other

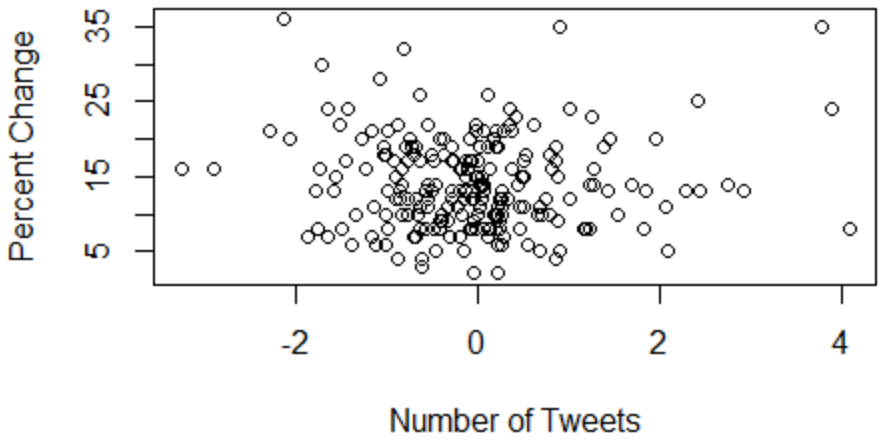

Figure 5.2: Number of tweets in a day versus the percent change of the Nasdaq

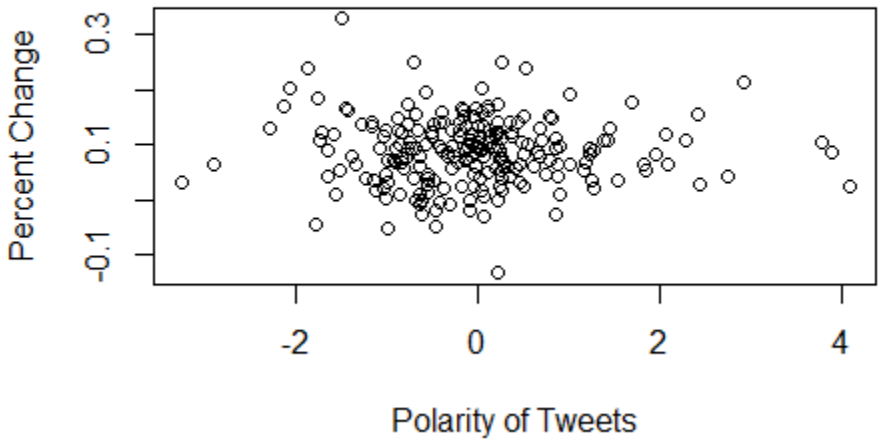

Figure 5.3: Polarity of tweets versus percent change of the Nasdaq

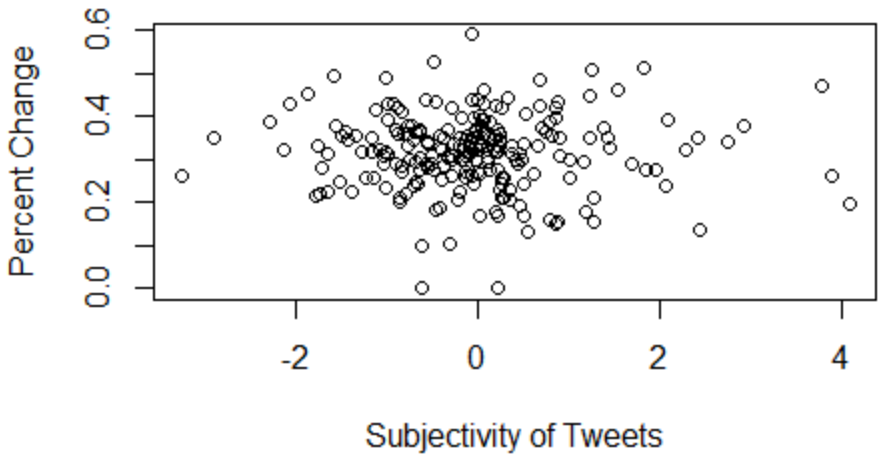

Figure 5.4: Subjectivity of the tweets versus the percent change of the Nasdaq

A simple linear regression model is used to try and determine if there is anything in the data of value. Unfortunately, this model seems to reveal that there is no correlation between our y and x variables. The null hypothesis (H_0) is that there is no correlation between the predictor variables and the response. The alternative (H_1) is that there is a correlation. Given an $\alpha = 0.05$, we fail to reject for each individual x variable as well as the model as a whole. The R^2 is 0.001 meaning that only .1 percent of the variance in the y is explained by the predictor variables. Clearly I am not getting what I am looking for here.

y = percent change of Nasdaq x1 = number of tweets

x2 = polarity of the days tweets x3 = subjectivity of the days tweets

```
Call:
lm(formula = y ~ x1 + x2 + x3)
Residuals:
    Min      1Q  Median      3Q     Max
-3.2382 -0.6488 -0.0064  0.4183  4.0705
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.082138   0.310886   0.264    0.792
x1          -0.001952   0.011893  -0.164    0.870
x2          -0.349819   1.237861  -0.283    0.778
```

```
x3          -0.230685   0.852669  -0.271    0.787

Residual standard error: 1.068 on 217 degrees of freedom
Multiple R-squared:  0.001153,   Adjusted R-squared:  -0.01266
F-statistic: 0.0835 on 3 and 217 DF,  p-value: 0.969
```

VI. Conclusions

The results of my experimentation give little information, other than what models are not useful. This is due to a mixture of data complexity, overly simplified models, and poor data collection. The sentiment data was not the best that could be chosen. In the future, a better script should be written to get tagged data that could filter out more of the useless information, then perhaps something of importance could be gleaned. The use of Textblob may be a poor choice since it is getting sentiment in an abstract sense. What is desired is sentiment as it relates to the market. It would be a good step forward to design sentiment analysis tool that weighted the value of significant words and phrases based upon what effect they have on the market.

In the future, it would also be good to try an ARIMA model with a shorter time span of data. This may allow the ARIMA model to pick up on a daily trend that may be able to predict with some accuracy the following few minutes. If ARIMA is implemented with other techniques or a better data set, the models could be extremely useful and insightful.

I have learned a great deal over the process of making this project and it has given me a lot to think about going forward. Much more time and effort will be required to make something that is more beneficial than simply as an exercise. People have made models that are informative, however, and I believe with many more hours and refinement, this could also become a valuable tool.

References

[1] R. Choudhry and K. Garg, “A Hybrid Machine Learning System for Stock Market Forecasting,” International Journal of Computer and Information Engineering, vol. 2, no. 3, 2008.

[2] A. Kazem, E. Sharifi, F. K. Hussain, M. Saberi, and O. K. Hussain, “Support vector regression with chaos-based firefly algorithm for stock market price forecasting,” Applied Soft Computing, pp. 947–958, 2013.

[3] A. Pak and P. Paroubek, “Twitter as a Corpus for Sentiment Analysis and Opinion Mining,” Universite de Paris-Sud, Laboratoire LIMSI-CNRS.

[4] J. Patel, S. Shah, P. Thakkar, and K. Kotecha, “Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques,” Expert Systems with Applications, vol. 42, pp. 259–268, 2015.

[5] S. Shen, H. Jiang, and T. Zhang, “Stock Market Forecasting Using Machine Learning Algorithms,” Stanford University.

[6] J.-H. Wang and J.-Y. Leu, “Stock Market Trend Prediction Using ARIMA-based Neural Networks,” National Taiwan Ocean University.