Predicting Stock Market Trends Using Sentiment Analysis of Twitter Data

Thomas Varley
Case Western Reserve University
Department of EECS
Cleveland, Ohio, USA
tjv32@case.edu

José Koluder-Ramirez
Case Western Reserve University
Department of EECS
Cleveland, Ohio, USA
jek158@case.edu

Long Pham
Case Western Reserve University
Department of EECS
Cleveland, Ohio, USA
ldp43@case.edu

***Abstract*—**Investing in the stock market has long been one of the most popular methods for individuals to attempt to increase their wealth. However, the stock market is extremely volatile, making its trends difficult to predict with high accuracy. The growing popularity of social media, especially Twitter, offers the opportunity to judge the public sentiment of different stocks. This paper will explore two new methods for predicting stock market trends using sentiment analysis of Twitter data. The explored methods generated sentiment scores for Tweets discussing different stocks in the S&P 500 over the previous seven days. Using two different APIs to collect tweets that included the name of a stock based on the time that they were tweeted, a gaussian naive bayes classifier and a neural network were used to classify whether the sentiment scores of tweets indicated that the stock price would increase or decrease on that day. Both the gaussian naive bayes classifier and the neural network were found to be more accurate than the null accuracy in predicting stock market trends, indicating that they are both valid prediction methods.

***Keywords*—**Twitter, Sentiment Analysis, Gaussian Naive Bayes Classifier, Stock Market, S&P 500

I. Introduction

As long as stock markets have been utilized as a means for a large corporations to publicly trade shares of ownership in their company, so to have individuals attempted to profit by predicting trends in the market. Over time, experts have employed numerous different techniques in order to attempt to predict these market trends to various degrees of success. Many of these prediction models are based off of Time Series Analysis, which utilizes previous market trends to predict the future success of different stocks. However, recent studies have proven Time Series Analysis to be an ineffective method for predicting trends in the stock market. Due to this, many recent studies attempting to predict future stock market trends have shifted focus to the Efficient Market Hypothesis, which encapsulates all relevant factors that affect public opinion of an individual stock in order to predict the future success of that stock [4].

It can be extremely difficult to accurately judge the public sentiment on any given topic due to a variety of underlying factors that can alter opinion form one individual to the next. However, the rise in popularity of social media offers an opportunity to witness changes in public sentiment in real-time. Over 3 billion individuals worldwide utilize social media every day to discuss their opinions on a variety of different topics, which in turn affects the public opinion surrounding that topic. Perhaps the most notable social media platform utilized in shaping public opinion is Twitter, which has over 126 million daily users. These users post 280-character tweets that concisely share their public opinion on topics ranging from sports to politics to the stock market. The sentiments shared in these tweets can be utilized to judge the overall public sentiment on any particular topic and make predictions based on that sentiment.

In this paper, we perform sentiment analysis on Twitter data in order to predict the future trends of individual stocks in the S&P 500. We accomplish this by assigning each individual tweet about a stock a positive (1) or negative (0) score and then calculating the mean value of the individual tweets to arrive at an overall sentiment score for that stock.

Since its founding in 2006, Twitter has demonstrated great usefulness as a tool for both public sentiment analysis and predicting real-world outcomes. Gerber [3] has utilized Twitter data to predict when, where, and what type of crime will occur in Chicago, Illinois. Odlum and Yoon have utilized Twitter to track the spread of the Ebola virus in the developed and developing world. Gayo-Avello [2] has utilized Twitter sentiment analysis to attempt to predict the outcomes of various elections. These studies all demonstrated the effectiveness of Twitter as a tool to predict future events in the real world.

Based on the previous success of other studies in utilizing Twitter data sentiment analysis to predict future outcomes, we can predict that this study will derive similar results. If we utilize a gaussian naive bayes classifier and a neural network in order to predict whether or not a stock will increase or decrease in price, then we anticipate both models being more accurate than the corresponding null accuracy of the data set.

The remainder of this paper will be structured as follows. Section II discusses previous works related to this study. Section III discusses the data set that was utilized in this study and how data was collected. Section IV discusses the analysis performed on the data set, including an explanation of the sentiment analysis and correlation analysis performed. Section V discusses the results of analysis. Section VI discusses conclusions that can be made based on the results of the analysis. Section VII discusses future work that can be done based on the the findings of this study in order to create a more accurate predictive model of stock market trends.

II. Related Work

One of the well-known studies we want to acknowledge is the work of Mankar et al [6]. In their paper, they utilized not only Twitter API to obtain daily tweets, but they also dug into Yahoo finance API to get the respective stock data. Through trying out a number of methods, Mankar's group concluded that Support Vector Machine (SVM) provides the most efficient and feasible model of prediction. Li, Zhou, and Liu [5] pointed out that with one hundred forty million tweets being generated every day, Twitter has become an excellent tool for them to conduct an analysis on public mood daily, therefore creating a correlation between public mood and stock price fluctuation. Acknowledging it's harder to analyze public with a broad range of different moods, Li's team split the public's feelings into six general categories: anger, disgust, sadness, happiness, surprise, and fear. Taking the frequency of each category, they derived the Pearson product-moment correlation coefficient, which represents the strength of the linear relationship between inputted variables. The result they collected shows that happy, sad, and anger appeared with higher frequency and have a stronger correlation with the stock result. Furthermore, sad-mood related vocabulary impacts the stock market more significantly than the other five. Lee and Paik [1] conducted a study by building a real-time processing system to analyze tweets for finding the correlation with the stock market: Associate with 77% accuracy of Twitter classification is 80% of separation of increase or decrease of stock value.

In our study, we acknowledged the tools and approaches these studies used and decided to take another route. Sharing a similar intention of creating a Neural Network, we also came up with a way of using Gaussian Naive Bayes Classifier with our approach. We then compared the accuracy value from both approaches with the null accuracy of the test data, and the utilization of Gaussian Naive Bayes Classifier returned a higher accuracy than the Neural Network. Not to say that Gaussian Naive Bayes guarantees to create higher accuracy correlation, but with the small sample that we used, it's safe to say that Gaussian does indeed have validity to it.

III. Data Set

In order to conduct analysis, recent Twitter data from the past weeks was obtained using two separate sources: Tweepy and GetOldTweets3.

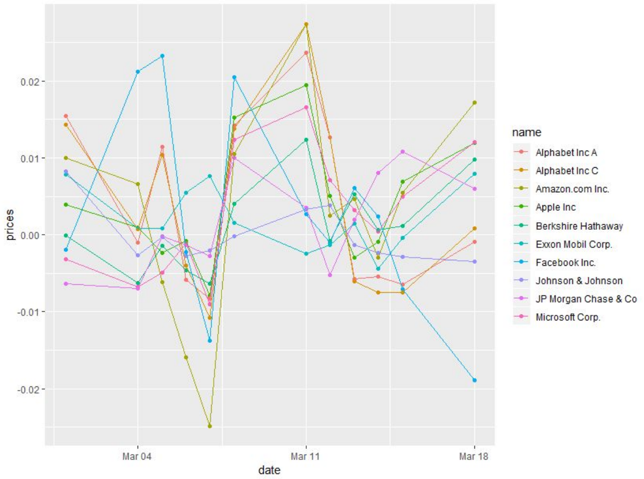
Tweepy, the official Twitter API, was used to obtain tweets between the dates of April 15, 2019 and April 18, 2019. The data set included all tweets obtained by Tweepy each day. Each company within the S&P 500 was utilized as a search parameter and each associated list of tweets was returned within a single day. Using the API IEXFinance, the associated change in stock price of the search stock was added to each entry row. There were approximately 2000 entries each consisting of a date, a stock name, the list of tweets with that stock name from the listed date, and a binary variable for direction of price change. The stock price was listed as either 0 or 1 indicate a decrease or increase in price, respectively.

Another API, GetOldTweets3, was used to collect tweets between the dates of March 1 and March 18. Data was collected for each day for each stock from the top ten stocks listed in the S&P 500. Additionally, the price change, listed as either 0 or 1 indicate a decrease or increase in price, were added to each entry row in the dataset. This totaled approximately 115 entries each consisting of a date, a stock name, the list of tweets with that stock name from the listed date, and a binary variable for direction of price change.

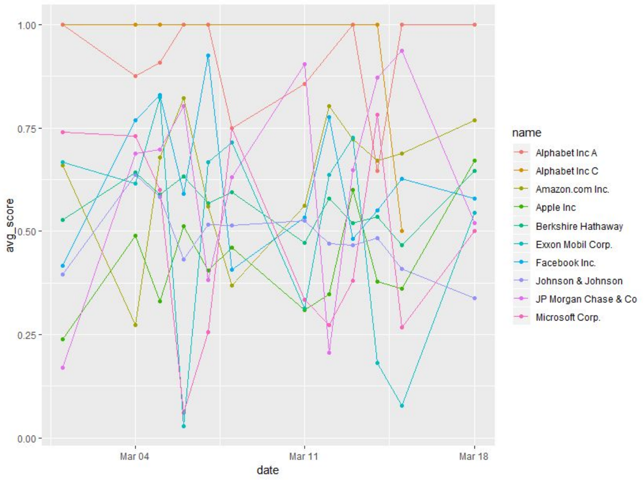
The two different datasets were created using two different APIs to collect the lists of tweets. The GetOldTweets3 API was able to retrieve tweets from earlier than 7 days. However, less tweets were returned than using Tweepy and often the tweets were simply links to news articles and not individual's feelings about certain stocks on certain days.

IV. Analysis

In order to create a model to analyze individual tweets as positive or negative, a dataset provided by Stanford University was used to fit a naive bayes multinomial classifier. The dataset included the text of each individual tweet and a corresponding score of either 0 or 4 to indicate whether the tweet referred to the stock in a positive or negative manner, respectfully. There was a total of 1.6 million labelled tweets. Each tweet was 'cleaned' by splitting contractions into separate words, taking out hyperlinks, removing unnecessary articles, making all letters lowercase, and removing emojis and unusual symbols. This cleaning was done to remove irregularities that may adversely affect the fitting of the naive bayes classifier. Each text entry was then vectorized using Term Frequency Inverse Document Frequency (TFIDF) which calculates the relative frequency of a word in a document. For a word, the higher the TFIDF frequency, the more rarely a word appears. Using 95% of the data, the classifier was fit to the data. The text of the tweets contained in this portion of the data were transformed using TFIDF and a document term frequency matrix was returned indicating how the relative inverse frequency of words contained in the training data set. A Multinomial Naive Bayes Classifier was then fit using the document term frequency matrix returned for each text of each tweet with the label being whether the tweet was positive or negative. The remaining 5% of the data was used to test the classifier. Each of the tweets contained in this test set were transformed into a document term frequency matrix and each matrix was the input feature to the Multinomial Naive Bayes Classifier. The label positive or negative was the target variable. An accuracy of approximately 80% was achieved in predicting whether a stock was positive or negative.



The above figure shows the prices of 10 sample companies chosen from the S & P 500 for the dataset taken from GetOldTweets3 for each date.



The above figure shows the average sentiment score for each date.

For both datasets, the multinomial naive bayes classifier was used to analyze the sentiments of tweets concerning stocks was applied to each individual tweet in each respective tweet list. The average sentiment score and standard deviation were recorded for each list of tweets recorded for each stock for each trading day.

To classify the direction of stock price for each individual stock on each trading day, a gaussian naive bayes classifier was fit to the data where the features were standard deviation alone, the average score alone, and average score and standard deviation together. The direction of the stock price was the target variable. The sklearn python library was used to create and fit the naive bayes classifier. Additionally, using the dataset found using the GetOldTweets3 API, a neural network was fit to the dataset to predict the target variable of stock price direction. The Keras library was used to create and fit the neural network.

Certain days for certain stocks in both datasets did not return any tweets or returned only a few tweets. To prevent a few tweets largely determining what the average score and standard deviation was for that day for that stock, the entire row of data was removed if the number of tweets was less than 400. This significantly reduced the number of entries that were included in the dataset provided by Tweepy to approximately 100 entries. However, each entry had a significant number of tweets to more accurately assess the average sentiment score.

To assess the relation between the average score, standard deviation, and number of tweets within a trading day for a certain stock versus the direction of the stock price, Gaussian Naive Bayes Classifiers were fit to different selections of features and then tested on a portion of the dataset. The accuracy of the classifier was compared to the null accuracy which was determined the greater of the percent of positive price change data points or the percent of negative price change data points depending on which was larger. This was used instead of 50% because we wanted to compare the fit classifier to a classifier that would predict positive or negative depending on which had a greater frequency.

The only feature that had a significant correlation with the direction of price change was the standard deviation of the sentiment scores. All other features had a correlation coefficient of approximately 0 while the standard deviation had a correlation coefficient of -0.2 indicating a weak negative correlation. Accordingly, the Gaussian Naive Bayes Classifier fit to the dataset provided by Tweepy had an accuracy of 62.5% compared to the null accuracy of the test data of 56.3%.

A neural network was used to predict the direction of stock price using the dataset provided by the GetOldTweets3 API. The Keras deep learning library was used to build the neural network. The feature provided was the standard deviation of the sentiment scores of the tweets. The target variable was the direction of the stock price which could only be positive or negative. 3 layers were used to create the neural network. The first layer had 12 neurons and received the input of the standard deviation of the sentiment scores. The next layer had 8 neurons. The final layer had one layer which was either 1 or 0 indicating a positive or negative stock price direction respectfully. The neural network was used to predict the direction of stock price for a test sample. The model had an accuracy of 53.9% compared to a null accuracy of 50.4%.,

V. Results and Discussion

Through the analysis of the sentiment of tweets and the direction of stock prices, we have found that the sentiment scores of tweets that contain a stock's name can be a meaningful predictor to the direction of the stock price. The Multinomial Naive Bayes Classifier, fit to the standard deviation of the sentiment scores of tweets provided by Tweepy, had an accuracy of 62.5% of predicting the direction of the stock price compared to a null accuracy of 56.3%. This indicates that the standard deviation of the sentiment scores is able to be used to predict the direction of the stock price. Additionally, the neural network that was trained using the standard deviation for the tweets provided by the GetOldTweets3 API had an accuracy of 53.9% when tested on another test set compared to a null accuracy of 50.4%. These models indicate that the standard deviation of tweet sentiment scores is a somewhat reliable predictor for the stock price direction for a day for an individual stock. In future work, we would like to include more data in order to improve our model.

VI. Conclusions

Through our analysis, it was determined that the standard deviation of the stock price was a useful predictor for the direction of the stock price. This indicates that the spread of the sentiment scores of tweets concerning certain companies is a better predictor than the average score of the sentiment scores. This negative correlation between the standard deviation of sentiment scores and the price direction as well as the accuracy of the models indicates that the larger the standard deviation the greater the change in stock price. This indicates that if the sentiment scores of tweets have a larger spread, the correlated stock price should have a greater chance of being negative. This implies that if there is more variability in what twitter users feel about stock prices, there is a weak correlation to the stock price decreasing.

VII. Future Works

There are various avenues that the methods and ideas discussed in this paper could be applied further to better and more accurately predict stock prices. One potential source for further study is to divide twitter sentiment into more than simply positive or negative tweets. The sentiment could be one of many levels of positivity or negativity. This division of emotions into more categories based upon the magnitude of positivity could allow a greater understanding of the magnitude of emotion within the tweet are more important for determining the direction of a stock's price. For example, the number of extremely negative tweets may be more important in determining the price direction of a stock than the tweets labelled as moderately negative. This would require a more sophisticated classification system and possibly a training set that included tweets labelled with different magnitudes of positivity. Additionally, for companies that receive many tweets, one could further investigate the hourly price changes and how they correlate to the sentiment of tweets sent within that hour.

Acknowledgements

The authors would like to thank Dr. Pan Li and Tianxi Ji for their continued instruction and help throughout the development of this project.

References

[1] C. Lee and I. Paik, "Stock market analysis from Twitter and news based on streaming big data infrastructure," *2017 IEEE 8th International Conference on Awareness Science and Technology (iCAST)*, Taichung, 2017, pp. 312-317. doi: 10.1109/ICAwST.2017.8256469, URL:http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8256469&isnumber=8256413

[2] Gavo-Avello, Daniel, “"I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper" -- A Balanced Survey on Election Prediction using Twitter Data”, *Social Science Computer Review,* Apr. 2012, URL: https://arxiv.org/abs/1204.6441

[3] Gerber, Matthew. “Predicting crime using Twitter and kernel density estimation” Decision Support Systems, 22 Feb. 2014, URL: https://www.sciencedirect.com/science/article/pii/S0167923614000268

[4] Kuepper, Justin. “Efficient Market Hypothesis (EMH) Definition.” *Investopedia*, Investopedia, 12 Mar. 2019, www.investopedia.com/terms/e/efficientmarkethypothesis.asp.

[5] Qian Li, Bing Zhou and Qingzhong Liu, "Can twitter posts predict stock behavior?: A study of stock market with twitter social emotion," *2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, Chengdu, 2016, pp. 359-364. doi: 10.1109/ICCCBDA.2016.7529584, URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7529584&isnumber=7529520

[6] T. Mankar, T. Hotchandani, M. Madhwani, A. Chidrawar and C. S. Lifna, "Stock Market Prediction based on Social Sentiments using Machine Learning," *2018 International Conference on Smart City and Emerging Technology (ICSCET)*, Mumbai, 2018, pp. 1-3. doi: 10.1109/ICSCET.2018.8537242, URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8537242&isnumber=8537238