CS591 Network and Markets **Final Project Report**short line

**Social Media Impact on Stock Market & Price**

Scarleth Estevez

Nathan Galloway

Sang-Joon Lee

Dec 16, 2016

# **Social Media’s Impact on Stock Market & Price**

A study of stock market behavior on news articles sentiments

Sang-Joon Lee, Nathan Galloway, Scarleth Estevez

Department of Computer Science

Boston University

Boston, MA

email: {sangjlee, nwg, scarlet} @bu.edu

## **Abstract**

In a world dominated by the internet and social media, we wanted to determine what effect social media played in the stock market. We are researching the effect of news article in social media of companies listed in stock market and whether or not the sentiment of articles and spread throughout the network is correlated with the performance of stock market prices. Previous research suggests that there is a correlation between news articles and stock price movements, and that there is lag between when the new articles are released and when the stock price moved to reflect the news. In our study, our hypothesis is that there will be a positive correlation between positive news articles and the stock market price and a negative correlation between the stock market prices and the number of negative news articles. We examined over 60,000 news articles on 26 different companies, across multiple sectors. Our analysis show that there is strong correlation between the volume of positive or negative news and stock market prices. We also found that there is lag of approximately 5.9 days before stock prices are affected by news articles.

## 

## **Keywords**

Social Media Analytics, Stock Market Prediction, Sentiment Analysis

## 

## **INTRODUCTION**

The process of determining the impact social media and the news has on the stock market began with searching for specific companies to serve as our case study. Over the course of our study, we examined over 60,000 news articles on 26 different companies, across multiple sectors from Fortune 500 companies. Originally, we were going to study both the news and Twitter, but Twitter’s API only allows the gathering of data from the previous 7 days, therefore we were unable to draw data over a long enough period of time to provide an appreciable amount of data. Our source for news articles was Webhose.io[[1]](#footnote-1), a news scraping service that provides the previous 30 days of news for free. Among their various filtering options, they include the overall sentiment of a news article. To achieve this, webhose.io[[2]](#footnote-2) uses Stanford CoreNLP[[3]](#footnote-3). This tool analyzes the adjectives assigned to the entity you are searching for, and weighs them as positive, negative, or neutral. We used this feature to track the daily changes in both the quantity of news about each company as well as the sentiment of each article. Since we were looking for daily trends, we used the closing stock values and the total number of daily articles. Our methods of analysis included finding the correlation between the stock prices and the number of news articles of each sentiment, and determining the Granger Causality of the news on the stock.

## **RELATED WORK**

Although there were a number of work on analyzing Twitter to predict future markets, there has not been considerable published research on studying correlation between news articles and stock market prices. Huberman and others [3] studied the effect Twitter has on box-office revenue for movies and shows that their predictor based on Twitter can outperform the market-based predictor. Bollen and others [1] shows that collective mood states derived from large-scale Twitter feeds are correlated to the value of the Dow Jones Industrial Average (DJIA) over time using technique such as granger causality, however, the study was based index market indicator rather but does not focus on individual stocks. Similarly, Qian and others [4] studied Twitter data to measure emotion and predict stock behavior. There were some research on correlation of news articles. Ming and others [5] studied the stock market prediction by using the daily article from The Wall Street Journal to predict the closing stock prices on the same day.

## **DATASETS**

We examined on 26 different companies, across multiple sectors selected from Fortune 500 companies as shown in Table I. We scrapped news articles for each of listed companies from Webhose.io, a data as a service platform over 30 days, from Nov 7, 2016 to Dec 6, 2016. We also gathered stock market price for each listed companies over same period of time.

|  |  |  |  |
| --- | --- | --- | --- |
| Amazon | Facebook | Johnson & Johnson | Time Warner |
| Apple | Ford | McDonalds | Toyota |
| Bank of America | GE | Microsoft | Twitter |
| Boeing | Goldman Sachs | Netflix | Verizon |
| Citigroup | Google | Nike | Walmart |
| CVS | Honeywell | Pfizer |  |
| Exxon | IBM | Starbucks |  |

Table I: Companies Analyzed

## **WEBHOSE.IO**

We gathered our data from Webhose.io, a data as a service (DaaS) platform, which provides an API that scraps online articles and returns results in JSON format. Webhose.io gathers live data from multiple of news sites, blogs, forums, and other online data sources[[4]](#footnote-4). Webhose.io API has named entity extraction capabilities which allows users to filter posts by specific entities such as organization. This feature reduces ambiguity in text; words that have multiple meanings are analyzed based on context. For example, a search query for organization Apple, will return news articles, forums and blogs that mentions ‘Apple’ as company. We used this for looking at the articles as well as looking at number of shares on the searched article over various social media, such as Twitter and Facebook.

The following is an example of JSON query result on organization Apple for positive sentiment online articles over last 30 Days. The query result is returned as JSON format where each post object contains different properties mapping string keys to values. It returns various data including the article text, published date, and number of social media shares.

|  |
| --- |
| "thread": {  "uuid": "2df2350d161f2dd484f3097cf5810ab103947e75",  "url": "http://omgili.com/ri/\_0JOtn.4SCoRuTimvW9\_C\_7IQx1rVk1r\_ZpsaQFmDeRiTFlMNqhjWuqzjzFX8PitAAmtS50AOtU-",  "site\_full": "bgr.com",  "site": "bgr.com",  "site\_section": "http://bgr.com",  "site\_categories": [  "news",  "tech"  ],  "section\_title": "BGR",  "title": "The new MacBook Pro is selling much better than you thought",  "title\_full": "The new MacBook Pro is selling much better than you thought",  "published": "2016-11-08T16:46:38.287+02:00",  "replies\_count": 0,  "participants\_count": 1,  "site\_type": "blogs",  "country": "US",  "spam\_score": 0.0,  "main\_image": null,  "performance\_score": 0,  "domain\_rank": 2702,  "social": {  "facebook": {  "likes": 6,  "comments": 0,  "shares": 6  },  "gplus": {  "shares": 0  },  "pinterest": {  "shares": 0  },  "linkedin": {  "shares": 1  },  "stumbledupon": {  "shares": 0  },  "vk": {  "shares": 0  }} |

Figure 1: Webhose.io JSON query of Apple with positive sentiment

## **STOCK MARKET PRICES**

During our study, we gathered stock closing prices from Yahoo Finance[[5]](#footnote-5) for the companies we analyzed. The stock prices were gather over 30 days, from Nov 7, 2016 to Dec 6, 2016 with same time period as data scrapped for news articles. With the stock market closed on weekends and Thanksgiving, we used the most recent closing price as the stock price for those days to interpolate the data.

## **DATASET CHARACTERISTICS**

We gathered our data for each of 26 companies from various sectors such as technology, consumer, manufacturing, healthcare, finance, energy. The number of companies in each sector were selected such that they are evenly distributed as possible. We had much difficulty finding companies in energy sector with enough volume of articles available for analysis.

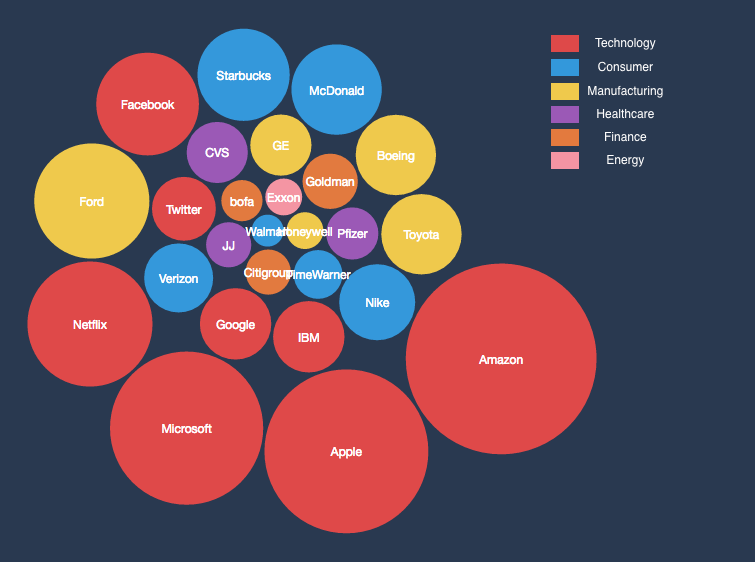


Figure 2: Volume of news articles for each company studied

The figure above shows the comparative volume of news articles gathered for each of the companies we studied, color-coded by each business sector. As you can observe from the visual representation, there were significantly large number of news article in technology sector than others over the 30 day period. Surprisingly, the financial sector showed lower than expected volume of articles.

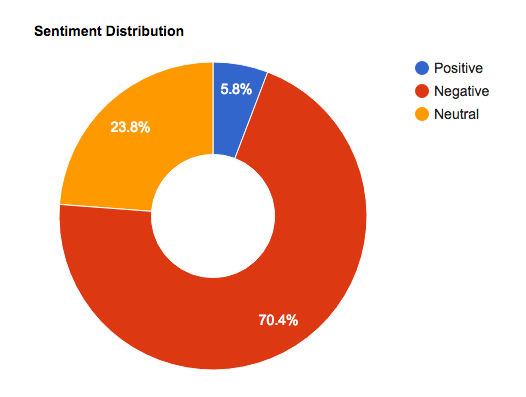


Figure 3: Breakdown of sentiment of the news articles

The figure above contains a proportional breakdown of the sentiment of the news articles scraped from Webhose.io (Positive, Negative and Neutral sentiments). We were surprised at how disproportionate the amounts were, with positive news only making up under 6% of the total articles.

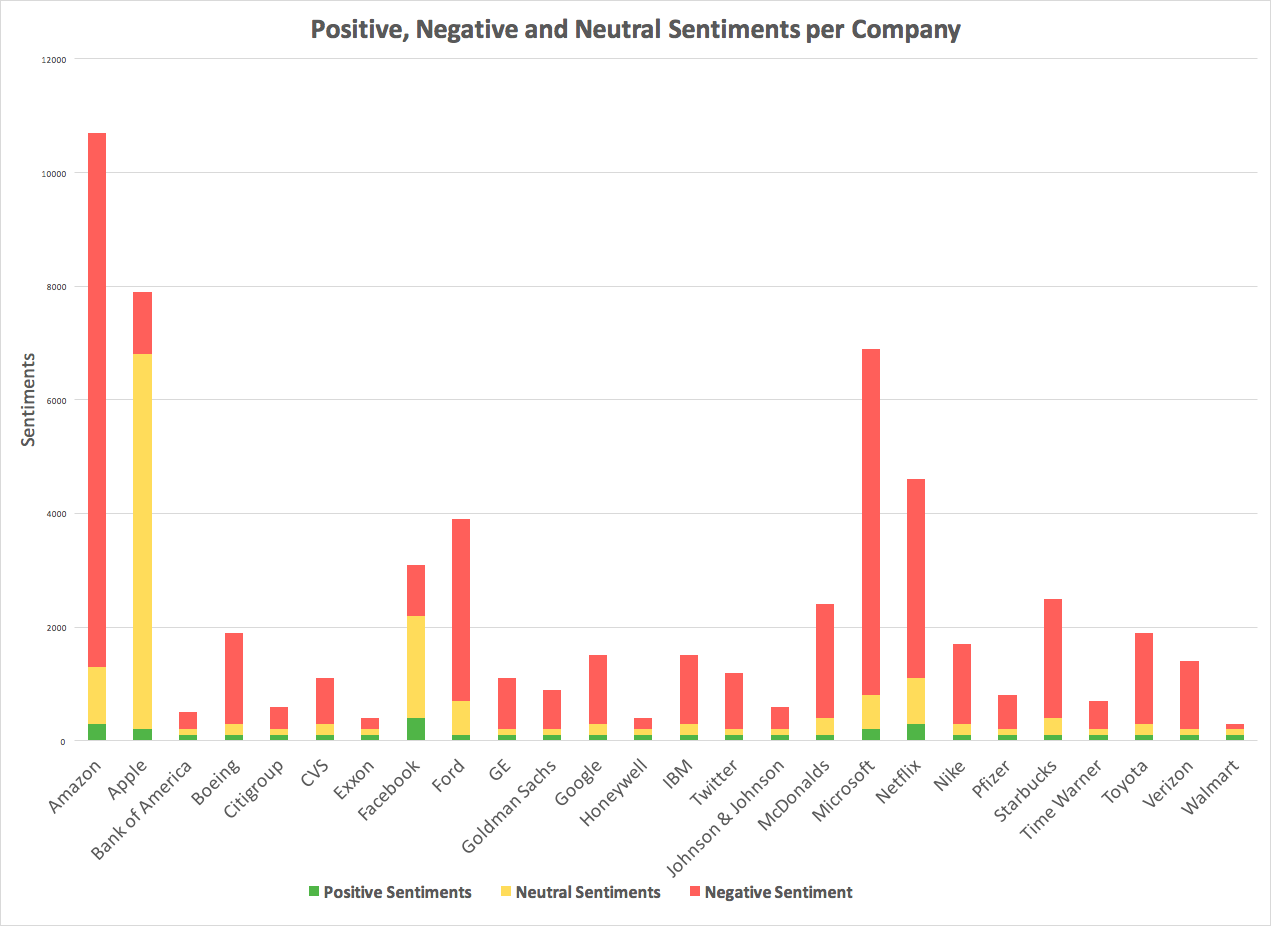


Figure 4: Volume of Positive, Negative and Neutral Sentiment articles per selected companies

The graph above shows the total amount amount of articles gathered for each company in our study. It is then further broken down by sentiments; positive, negative and neutral, corresponding to the colors. In almost all cases, the number of negative articles makes up over half of the news scraped for each company.

## **METHODOLOGY**

As shown in the figure below, upon choosing a company for our study, we queried Webhose to gather all of the news articles about that company over last 30 days. We then took those results and separated them by sentiment before further dividing them by day. The result of this was an array of each sentiment, with a daily count of articles for that sentiment. We then gathered the publicly available stock data for the same period of time and used them in conjunction with our arrays to find the correlation coefficients listed below. After finding the direct correlation, we then began searching to see if there was a cause-effect relationship between news and the stocks. To accomplish this, we used the Granger Causality algorithm, described below.

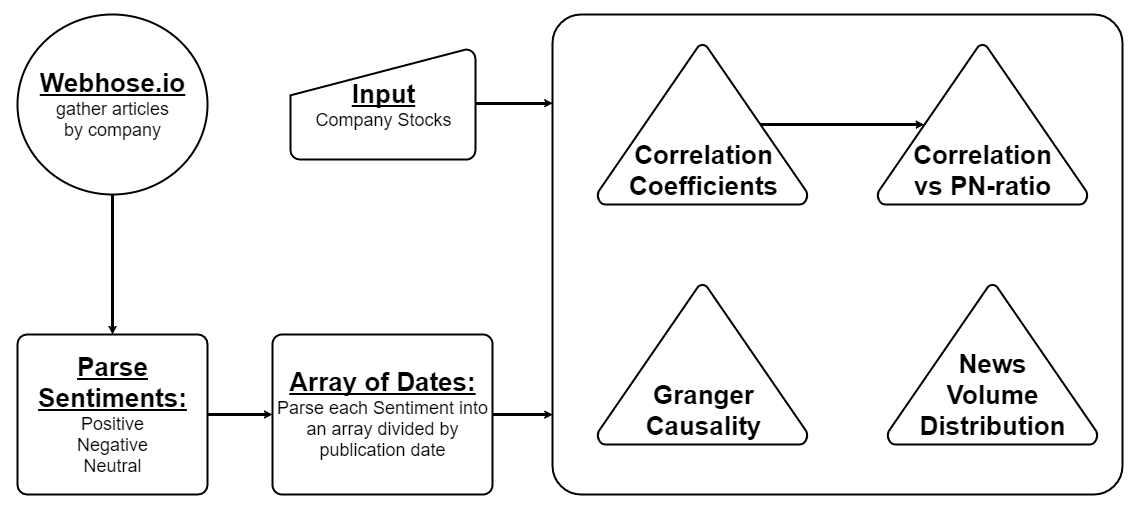


Figure 5: Methodology Overview

## **Correlation Analysis**

We investigated how closely linked the following datasets are with each other by calculating cross correlation of the following:

1. Volume of Positive articles and Stock Prices movement on the same day
2. Volume of Negative articles and Stock Prices movement on the same day
3. Volume of Positive articles and Stock Volume on the same day
4. Volume of Negative articles and Stock Volume on the same day

To quantify the sentiments for a company news article, we measured the ratio of positive to negative new article volume per day, similar to research method used by Huberman and others [3]. The positive to negative ratio, PN Ratio is defined as shown in equation below.

This is under assumption that a company with more positive sentiment news article than negative news article would perform better (i.e. positively) in stock market. We calculated the correlation between PN Ratio and stock market price to see if there were any correlation in performance.

## **Granger Causality**

We analyzed two events over a period of time and determined if the first event can predict what the second event will do in the immediate future, similar to research method used by Bollen and others [1]. For example, if there is a large increase in positive news articles with stock market price rise the following day.

The Granger Causality uses a lag value, in our case days, to give a null hypothesis of the causality. It analyzed the number of articles of a given sentiment, gave us a p-value corresponding to the likelihood of the news NOT predicting a change in the stock price. Anywhere the p-value was below .05, the null hypothesis failed, giving us a 95% confidence of the news predicting the stock price. As shown in our results, our data revealed a lag of about 5.9 days before the news had an effect of the stock market.

## **RESULTS**

## **Correlation Results**

The below graph is a plot of the positive and negative sentiments correlated to the stock volume. All points to the right of the diagonal correlate more strongly with a positive sentiment, while all points to the left correlate more strongly with a negative sentiment.

## ../../Downloads/volume%20scatter.png

Figure 6: Positive Sentiment and Volume Correlation vs. Negative Sentiment Volume Correlation

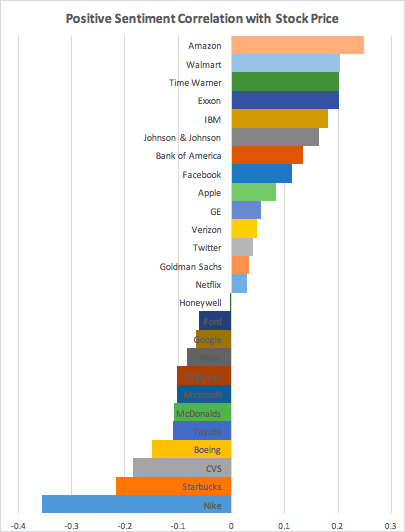
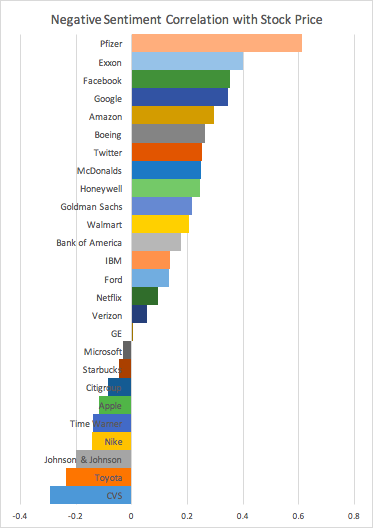
 

Figure 7: Positive & Negative Sentiment Stock Price Correlation

The above figures show the correlation between positive and negative sentiments and the stock price for each company. Given the overall trends found in the stock market, many companies showed only a weak correlation.

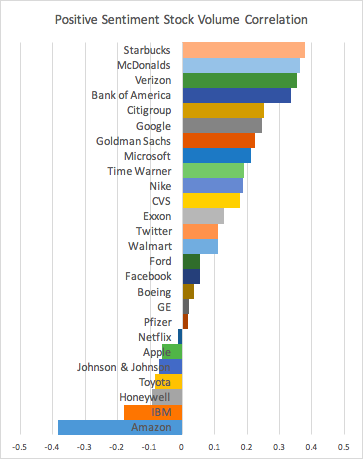
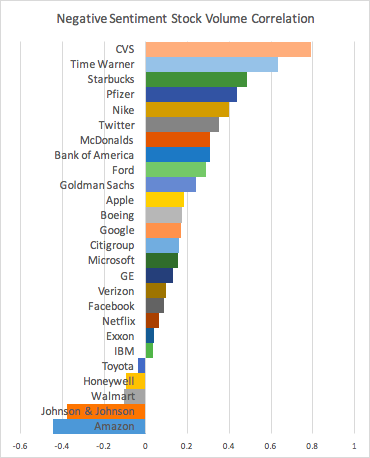
 

Figure 8: Positive & Negative Sentiment Stock Volume Correlation

Figure 8 shows the correlation between sentiments and the stock volume for each company. By using the volume instead of price, it helps to show a more accurate correlation between the sentiments and how much stock is being traded on the market. We observed that approximately ⅔ of companies has strong correlation between news articles that mentions their company name and volume of stock trades made during that day.

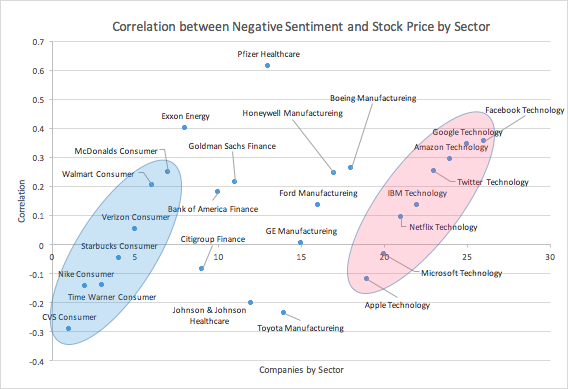
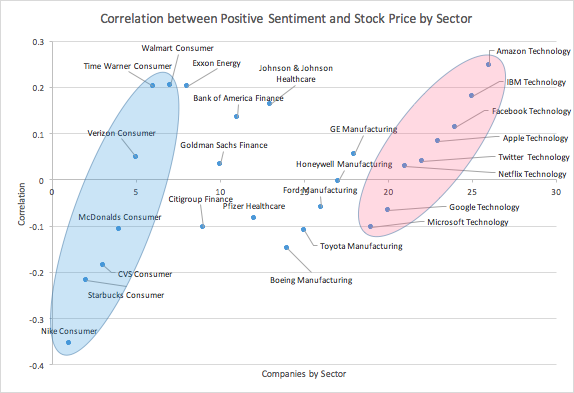


Figure 9: Correlation between Sentiment and Stock Price by Sector - Positive Sentiment (Left) and Negative Sentiment (Right) where technology sector highlighted in red ellipse has higher correlation on average and more clustered.

Figure 9 shows the correlation between sentiment and stock price group by different sectors. We can observe that form visual observation, the companies in technology sector (highlighted in red) has slightly higher average correlation with stock price than the companies in consumer sectors (highlighted in blue). Also, the companies in technology sector appears to be more clustered together in both Positive and Negative sentiments, implying that the stock prices correlate more in the similar direction than other sectors such as manufacturing or healthcare. One of the reason this behavior is observed may be due to more data samples scraped for technology and any other sectors, which may yield more accurate result. In addition, technology sectors are known to be more volatile than other sectors due to the longevity and the fundamental nature of the business.

## **Granger Causality Results**

The Granger Causality test was performed on each of the companies for Positive, Negative and Neutral sentiments. Table II below shows results for each of three sentiments for list of companies. The cell highlighted in green are p-values less than 0.05 (5%) which proves FALSE on the null-hypothesis (where null hypothesis is that the news does NOT cause a change in the stock price), which implis that there is high likelihood that the news DOES cause a change in the stock market price. Our results show that there is more likelihood of causal effect for negative and neutral sentiments than positive sentiments. Our study of the Granger Causality of news articles on the stock market revealed that the news impacted the stock market after a little over 5.9 days.

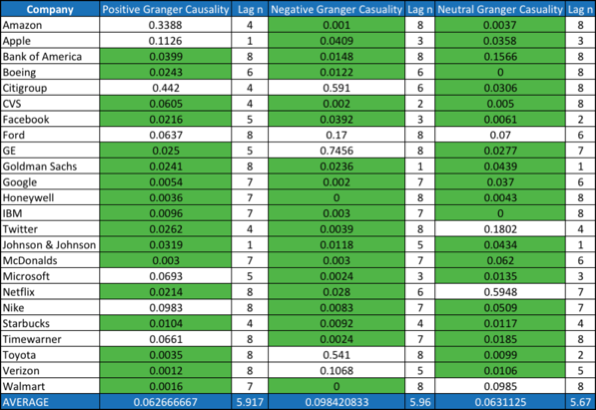


Table II: Granger causality relationship between volume of positive, negative, neutral sentiment articles and stock market prices

## **CONCLUSION & FUTURE WORK**

In this study, our analyses of sentiment to stock volume revealed a general positive correlation between the two. Our study of the Granger Causality of news articles on the stock market revealed that the news impacted the stock market after a little over 5.9 days. This result fell in line with related article, “Twitter mood predicts the stock market”[1], in which they found a 3 day lag between social media from Twitter and the stock market.

One of the difficulties we encountered was the fact that most large companies have a slow, upward trend in the stocks market, which skewed our results. This was why we began looking at the correlation between the articles and the volume of stocks being traded. While it wasn’t the perfect solution, we hoped it would make for a more normalized result to see how the news correlated to the change in volume of stock trades, rather than trying to counter the slow rise of stocks over time. The correlations between sentiment and stock price didn’t follow any of the expected trends, since the slow rise in stock price meant that even negative news articles correlated to positive stock growth. Using stock volume minimized those discrepancies for us. At the poster presentation, Prof. Byers discussed with us how we could have also looked at the sector growth averages to help normalize the stock data.

Another challenge was the spread of news articles available to us. Webhose performed all of its own analysis behind the scenes, so we were unable to provide any input in how it classified the sentiments. It follows a binary system rather than a graduated approach. If we had been able to design an algorithm that would give a weight to how positive or negative an article was, it could have helped to refine our results. Also, with only the previous 30 days available for free, we were limited in the scope of our research. There was also an extremely large range in the amount of news articles about the companies we were studying. Many of the companies were around 1,000 or 2,000 articles, Walmart only had 79 news articles over a 30-day period whereas Amazon had over 10,000.

Given the time and budget, we would have loved to extend our research to cover a longer period of time. Following news and stock trends over a 1-year period would have given us much more accurate results and better insight into the topic. We would also include Twitter data into our study and see how the spread and diffusion of the sentiment compares to what we gathered in the news. Since you can only gather the previous 7 days of data from the Twitter API, it would take several months of scraping data each day to get enough from Twitter to make it worth studying. We would love to validate our current processes by analyzing Twitter’s effect on the stock market to reach the same 3 day lag in Granger Causality that was found in the related work.

## **ACKNOWLEDGEMENT**

A special thanks to Professor Byers for his guidance in our research and suggestions on how to take our study further.

## **APPENDIX I**

All scraped data from Webhose.io and source code developed for analysis are shared on github respository for anyone to explore the data and analysis scripts. The following is link to the project github repository:

* <https://github.com/cs591B1-Project/Social-Media-Impact-on-Stock-Market-and-Price>

## **REFERENCES**

1. Bollen Johan, Huina Mao, and Xiaojun Zeng, “Twitter mood predicts the stock market”, Journal of Computational Science 2.1 (2011):1-8
2. C Castillo, M Mendoza, Barbara Poblete (2011), Information credibility on Twitter, Proceedings of the 20th international conference on World Wide Web., page 675-684
3. Asur Sitaram, and Bernardo A. Huberman. “Predicting the future with social media.” Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE WIC/ACM International Conference on. Vol. 1. IEEE, 2010.
4. Li, Qian, Bing Zhou, and Qingzhong Liu. "Can twitter posts predict stock behavior?: A study of stock market with twitter social emotion." Cloud Computing and Big Data Analysis (ICCCBDA), 2016 IEEE International Conference on. IEEE, 2016.
5. Ming, Felix, et al. "Stock market prediction from WSJ: text mining via sparse matrix factorization." 2014 IEEE International Conference on Data Mining. IEEE, 2014.

1. https://webhose.io/ [↑](#footnote-ref-1)
2. http://stanfordnlp.github.io/CoreNLP/ [↑](#footnote-ref-2)
3. http://www.programmableweb.com/news/webhose.io-api-now-features-named-entity-extraction-capabilities/brief/2015/04/02 [↑](#footnote-ref-3)
4. http://www.programmableweb.com/news/webhose.io-api-now-features-named-entity-extraction-capabilities/brief/2015/04/02 [↑](#footnote-ref-4)
5. https://finance.yahoo.com/ [↑](#footnote-ref-5)