Predicting Stock Market Fluctuations from Twitter

An analysis of the predictive powers of real-time social media

Sang Chung & Sandy Liu

Stat 157

Professor ALdous

Dec 12, 2011

1. Introduction

Advances in technology accelerate at ever-increasing rates. With the rise of new technologies in the field of the internet and social media, the popularity and importance of numerous social media platforms has risen to new levels, as more people spend more time online and companies follow their potential customers. One such social media platform that has seen an explosive rise in popularity is Twitter. Twitter is a real-time information network that connects its users to the latest information about subjects interesting to them. To do so, all users need to do is "follow" others in the field—whether they be experts, celebrities, or companies—to receive instant updates on their posts. The central idea of Twitter is the "tweet." Each post on twitter is called a tweet, and it contains a maximum of 140 characters. This format forces users to be concise, and thus each tweet is a short burst of condensed information that makes it easy to read, easy to follow, and—for our purposes—easy to statistically data mine useful information about societal trends.

Twitter has seen explosive growth over the past few years. Now up to 200 million registered users, Twitter sees 50 million users a day and 400 million visitors a month. Approximately 1 billion tweets are generated by Twitter users every five days. The obvious popularity of Twitter has led many people and celebrities to join Twitter in order to potentially connect with others and increase their popularity and awareness. With so many people tweeting about their various opinions about subjects ranging from toothpaste to the newest Apple products, Twitter is a rich source of real-time information regarding current societal trends and opinions.

Behavioral economics tell us that people are not rational consumers and individual behaviors and decisions are greatly affected by emotions—and indeed by the opinions of others. This should hold true for societies at large; that is, society can experience mood states that affect

their collective decision-making. So, if each tweet is a condensed summary of a person's mood or opinion about a certain subject, then the aggregate of tweets about the subject should express the collective mood. By extension, public mood should be correlated with or even predictive of economic indicators.

This study attempts to examine Twitter's predictive potential of consumer purchasing by observing the relationship between societal Twitter trends in the technology sector and hourly stock prices of the top gainers and top losers of ten companies in the technology sector. We hypothesize that the trending mood in Twitter about the top gainers in the technology sector will be positive, while the trending mood about top losers will be significantly more negative compared to a baseline measurement of the trending mood in the overall technology sector.

2. Data

The data used in this study is collected from three different sources.

2.1.1. Twitter

Twitter's API can return data on the latest tweets in XML format. We conducted three separate searches on Twitter to return the dates and latest tweet content on 1) the top gainers in the technology sector, 2) the top losers in the technology sector, and 3) Na general search on technology and stocks. The searches were conducted using the AND/OR and quotations style Boolean search method, as well as using subject (hashtag) searches and tweets to and from certain users (to:company; from:company) on Twitter. The dates and contents of each Twitter post for these searches were obtained for the time period of November 29, 2011 to December 2, 2011.

2.1.2. Hu & Liu's List for Sentiment Analysis

To quantify the data collected from Twitter.com, we carried out what is called sentiment analysis. As a part of our opinion mining process, we wanted to have a quantitative way of measuring positive or negative sentiment of the selected Twitter community of our interest. We chose to use the sentiment list put together by leading researchers of this, Minqing Hu and Bing Liu. Commonly known as Hu and Liu's sentiment list contains about 6800 words that reflect either positive or negative sentiment. We compared each tweet to the list, and counted positive words with positive scores, and negative words with negative scores. In theory, tweets with more positive words than negative words would reflect a positive score, and vice versa.

2.1.3. Google Intraday Stock Prices

The highest and lowest gainers in stock were identified from Google stock prices as follows:

Top Gainers-Company and Stock Symbol	Top Losers-Company and Stock Symbol
1. Advanced Analogic Technologies-AATL	1. Omnivision-OVTI
2. Intevac, Inc.-IVAC	2. Mediware-MEDW
3. Trina Solar-TSL	3. Sapiens-SPNS
4. Canadian Solar-CSIQ	4. BMC-BMC
5. Exide-XIDE	5. Micron Technologies-MU

Actual intraday stock prices on these companies in the technology sector were gathered from Google Intraday Stock Prices using Volumedigger. The data is stock price fluctuation by the minute, which we reorganized into hourly stock price data by averaging the price across the hour. In addition, we standardized the stock price data for

each different type of stock in order to compare all stocks to one another. This data is also from November 29, 2011 to December 2, 2011.

3. Statistical Methods

3.1. Graphical Methods

Table 1. Comparison of aggregate sentiment scores of the highest stock gainers and losers against the baseline.

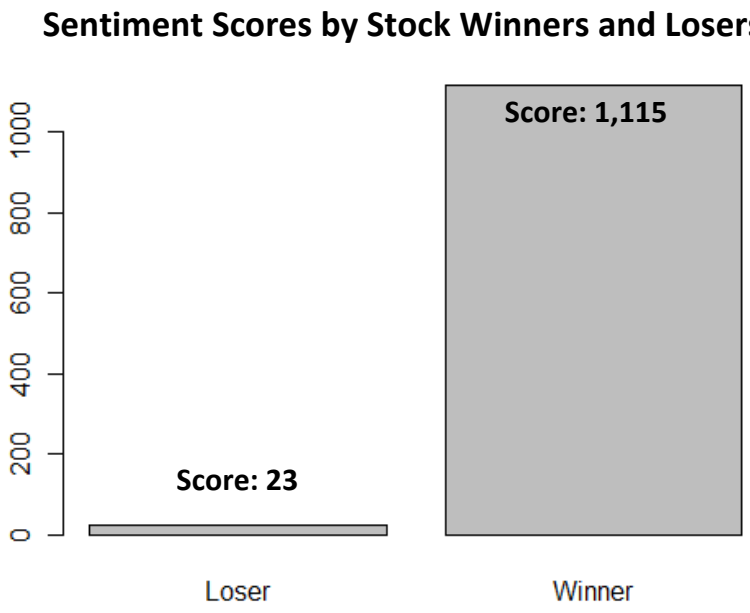
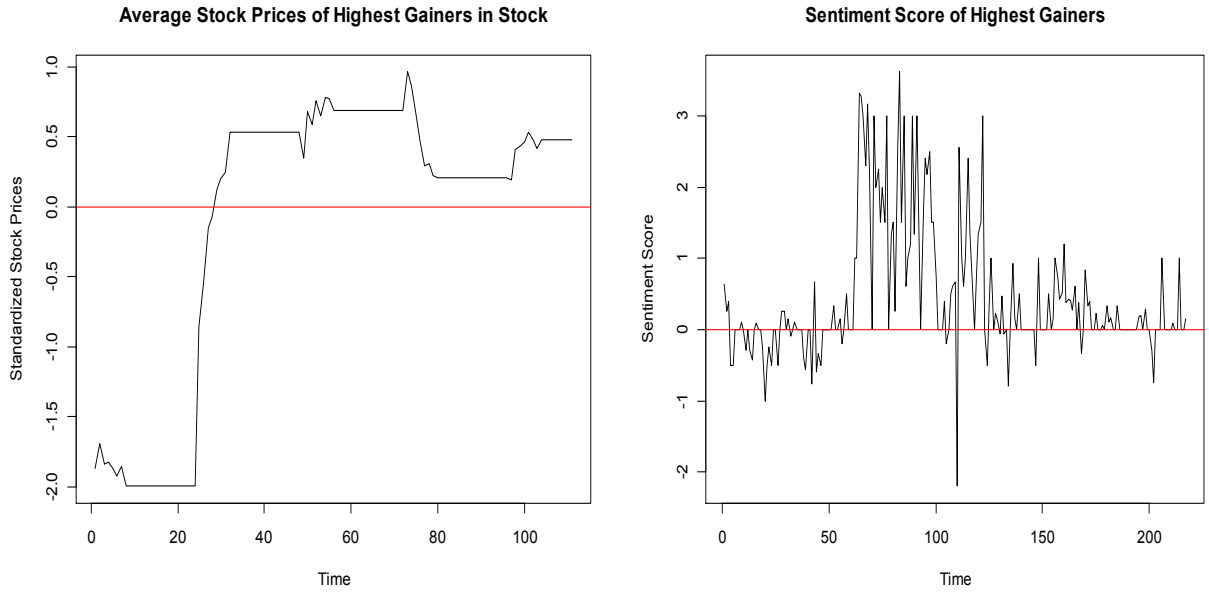


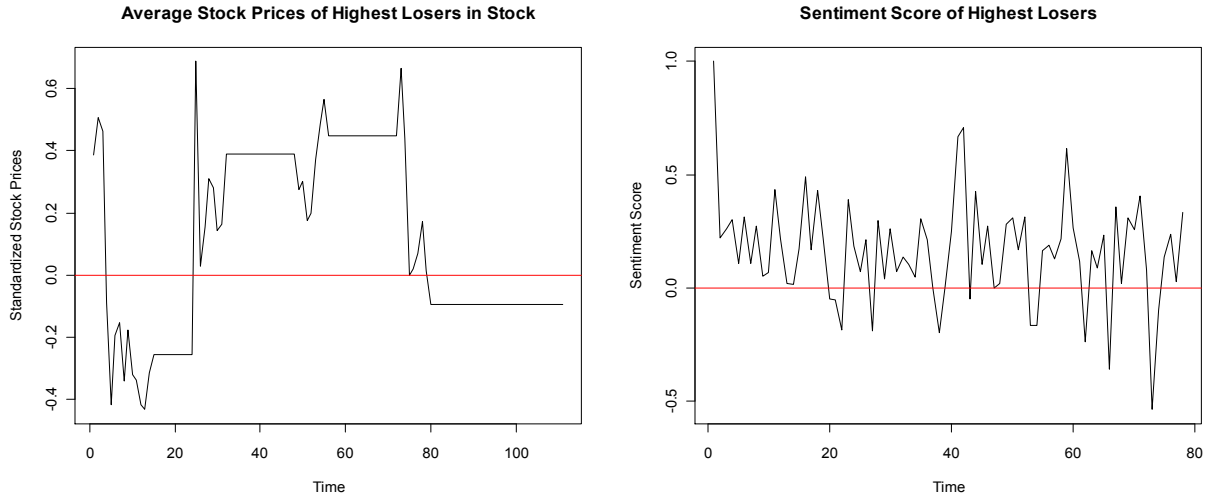
Table 1 shows the aggregate sum of sentiment scores for tweets in the highest gainers and highest losers, both differenced by the aggregate sum of sentiment scores for tweets in the overall technology sector. This shows the general positive or negative sentiments about the companies in question. Obvious from the barplot is that the aggregate sentiment score for the highest winners are much higher than the sentiment score for the losers, showing that there is a much more positive trending Twitter mood about the highest gainers in stock in comparison to the highest losers in stock.

Figure 2. Time plot of the stock prices and sentiment scores of highest gainers in stock



As we can see from Figure 2, the sentiment scores from Twitter seem to follow the average stock price, since the large upward spike in the stock prices at time index 25 is reflected at around time index 60 of the sentiment scores. In addition, the sentiment score seem to be mostly positive (above zero), while for the most part the stock prices are above zero as well (which is the mean of the stock since they are standardized).

Figure 3. Time plot of the stock prices and sentiment scores of highest losers in stock



However, in Figure 3, we do not see a similar pattern occurring for the highest losers in stock. There seems to be a slight linear downward trend over time of the sentiment scores, while the average stock price jumps to a high point and then decreases over time.

There are two points to be made from the analysis of these plots. First is that there are several noticeable areas of the stock price time plot that are flat and show no variation and show a sudden increase or decrease afterwards, and these occur at intervals. This can be explained by the way the stock market works. The stock market opens every weekday at 9:00am EST and closes every weekday at 4:00pm EST, and during its closed hours we have no data on hourly stock price fluctuations. So, for our study we took the closing price at 4:00pm every day and extended that price for the duration that the stock market is closed. This accounts for the flat areas of the time plot. Also, even though the stock market is closed, people can still put in buy or sell orders during the closed hours, which a computer program takes into consideration and computes an opening price for the next day—which can be very different from the closing price of the previous day. This accounts for the sudden upward or downward movements right after the flat, invariant hours of the closed hours of the stock market.

Furthermore, a comparison between the stock price plots of both the highest gainers and highest losers reveal that both see a significant positive upward spike around time 25, which occurred at 9:00AM on November 30. Therefore there may be some factor that has affected the stock prices in the technology sector across the board. Further research into this phenomenon indicates that this sudden positive movement is due to a join decision reached by the central banks of Europe, the US and other economic powers, which gave banks

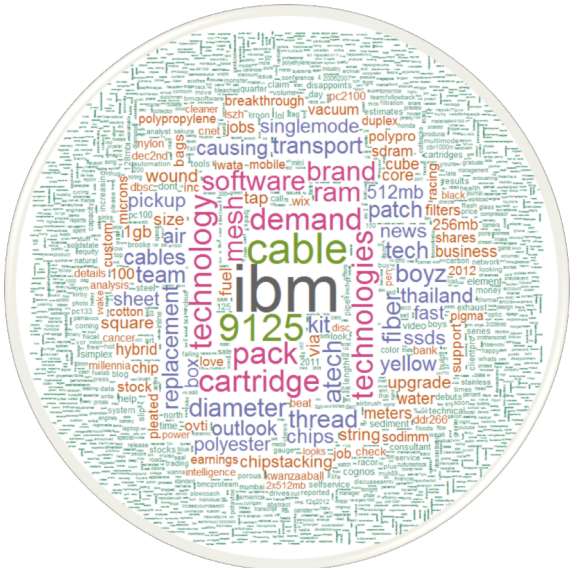
cheaper access to loans and US dollars. This was an effort to divert another credit crisis and improve the financial market, and had clear and immediate effects on the stock market prices.

Finally, these last three figures attempt to find the most frequently used words in the aggregate twitter posts, excluding the keywords used to generate the search results in the first place.

Figure 4. Word cloud of stock in the overall technology sector.

Figure 5. Word cloud of highest winners in stock in the technology sector.

Figure 6. World cloud of the highest losers in stock in the technology sector.

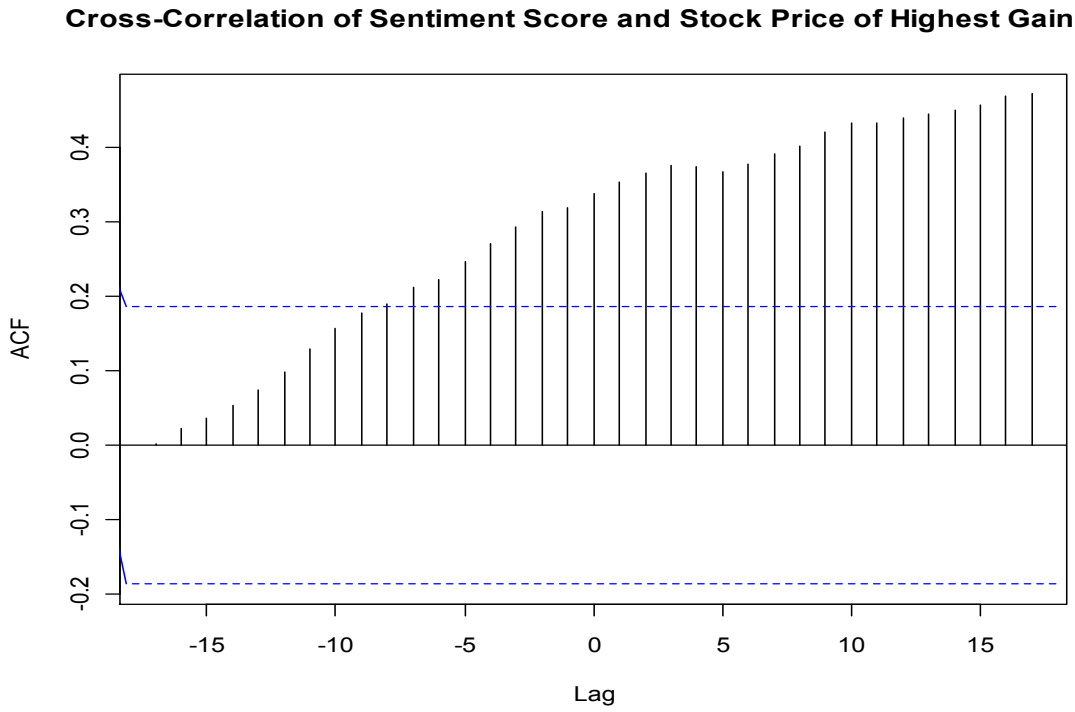


These three word clouds emphasize the most frequently used words of all the tweets. As we can see from Figure 5, the companies that are the highest gainers in stock are companies that emphasize green technology and renewable energy such as solar power. This shows that the companies with highest rising stocks are companies with an environmental focus, which is a societal trend at the moment as the threat of global warming becomes ever more threatening. However, there is no identifiable technology subsector in the highest losers as there is in the highest gainers, which we can see from further research on the companies comprising the highest losers—they are varied and different, involving imaging technology as well as insurance software. So, renewable, green energy seems to be the strongest gaining stock, and may be worthwhile to invest in.

3.2. Cross-Correlation

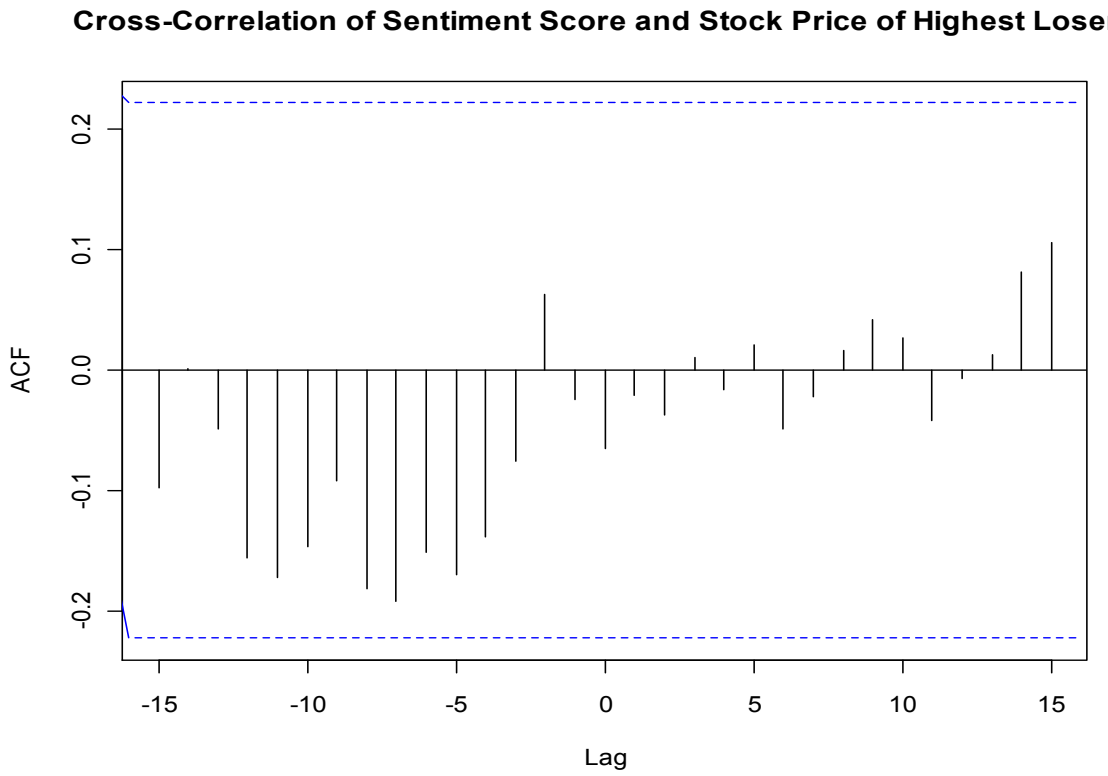
To see if the Twitter sentiment score is correlated and possibly predictive of stock prices, we decided to graph the cross-correlation between the two data sets. The CCF for the highest winners and losers in stock in the technology sector is displayed below.

Figure 7. CCF between the Twitter sentiment score and the stock price of highest gainers



As we can see from Figure 7, the Tweet sentiment score precedes stock price movement starting from about 7 hours beforehand for the highest gainers in stock. As time goes on, the correlation between tweets score and stock prices increases linearly. This shows that to some degree, tweet sentiments precede stock prices. However, tweets after stock price changes are even more strongly related with stock prices, meaning that after stock prices movements change, tweets are more likely to change in the same direction. However, strong cross-correlation is not observed in the CCF of the highest losers' stock prices and their sentiment score.

Figure 8. CCF of sentiment score and stock price of the highest losers in stock in the tech sector.



There is no significant correlation at any lag. However, there seems to be slight correlation (not significant) at around 5-10 hours beforehand, indicating there is slight evidence that the Twitter sentiment score for the highest losers may cause stock to decrease. The evidence for this is not significant. Furthermore, after lag 0, we do not observe the same increasing cross-correlation between stock price and twitter sentiment score. Intuitively, this makes sense. People are less likely to brag about their losses than their gains, which is why we see a slight relationship between twitter sentiment and stock price, but the stock price does not seem to negatively affect the twitter sentiment score, since people are less likely to tweet about their losses publicly.

4. Conclusion

We conclude that the twitter sentiment score may predict the movement of stocks if the sentiment score is trending positive, not negative. However, stock price movements are more strongly predictive of twitter sentiment movements. There is no significant predictive power of trending negative sentiment scores on stocks relating to the subject.

However, there were several limits to this study that we must address. We were limited to four days of Twitter data due to the constraints of Twitter's privacy policy. In addition, stock price fluctuations are unavailable during the stock market's closed hours while Twitter postings can be generated during any time of the day. Having access to stock price fluctuations even during the closed hours may provide a more accurate gauge of the effect of twitter sentiments on stock prices. The last point of interest is the methodology of sentiment analysis. There is an inherent problem with merely using a list of words with traditionally positive or negative connotations to score a body of text—this method cannot identify non-literal phrases such hyperboles and sarcasm, and may misconstrue the true sentiment of the text. In addition, Twitter posts can contain hyperlinks to images and other websites, which we did not address in this study and may or may not affect our outcome. Lastly, the sentiment lists are lacking in sector-specific terms that may have positive or negative connotations, which may also affect the sentiment scores of the Twitter data. These limitations illustrate the restrictions of our methodology and data, and suggest room for improvement in future studies.

5. References

Bing Liu (2010). “Sentiment Analysis and subjectivity”. *Handbook of Natural Language Processing, Second Edition, (editors: N. Indurkhya and F.J. Damerau), 2010.*

Bollen, Johan, Huina Mao, and Xiao-Jun Zeng. "Twitter mood predicts the stock market." *Journal of Computational Science*. (2011): 1-8. Print.

Breen, Jeffrey. "mining Twitter for consumer attitudes towards airlines." *Things I tend to Forget*. Wordpress, 04 Jul 2011. Web. 12 Dec. 2011. <http://jeffreybreen.wordpress.com/tag/sentiment-analysis/>.