Data Mining in Social Media for Stock Market Prediction

by

Feifei Xu

Submitted in partial fulfilment of the requirements
for the degree of Master of Electronic Commerce

at

Dalhousie University
Halifax, Nova Scotia
August 2012

DALHOUSIE UNIVERSITY

FACULTY OF COMPUTER SCIENCE

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled "Data Mining in Social Media for Stock Market Prediction" by Feifei Xu in partial fulfilment of the requirements for the degree of Master of Electronic Commerce.

Dated: August 9th 2012

Supervisor: _____

Readers: _____

_____

DALHOUSIE UNIVERSITY

DATE: August 9th 2012

AUTHOR: Feifei Xu

TITLE: Data Mining in Social Media for Stock Market Prediction

DEPARTMENT OR SCHOOL: Faculty of Computer Science

DEGREE: MEC CONVOCATION: October YEAR: 2012

_____

Signature of Author

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

In this thesis, machine learning algorithms are used in NLP to get the public sentiment on individual stocks from social media in order to study its relationship with the stock price change. The NLP approach of sentiment detection is a two-stage process by implementing Neutral v.s. Polarized sentiment detection before Positive v.s. Negative sentiment detection, and SVMs are proved to be the best classifiers with the overall accuracy rates of 71.84% and 74.3%, respectively. It is discovered that users' activity on StockTwits overnight significantly positively correlates to the stock trading volume the next business day. The collective sentiments for afterhours have powerful prediction on the change of stock price for the next day in 9 out of 15 stocks studied by using the Granger Causality test; and the overall accuracy rate of predicting the up and down movement of stocks by using the collective sentiments is 58.9%.

LIST OF ABBREVIATIONS USED

UGC	User Generated Content
NLP	Natural Language Processing
DJIA	Dow Jones Industrial Average
S&P 500	Standard & Poor's 500
WSJ	Wall Street Journal
SVI	Search volume index
EMH	Efficient Market Hypothesis
WOM	Word-of-mouth
CNG	Character N-grams
SVM	Support Vector Machines
SMO	Sequential Minimal Optimization
NB	Naïve Bayes
ME	Maximum Entropy
BOL	Bill of Lading
DT	Decision Tree
POS	Part-of-speech
GNU	General Public License
GC	Granger Causality
ADF	Augmented Dickey-Fuller

ACKNOWLEDGEMENTS

First and foremost I would like to thank Dr. Vlado Keselj, my supervisor for this thesis. His devoted effort in guiding me throughout not only this thesis but also the preliminary work has greatly benefited the process. I also would like to thank Dr. Vladimir Lucic from Barclays Capital for providing an internal report relevant to my thesis and confirming the significance of my research. Many thanks to Dr. Malcolm Heywood on my thesis committee and the DNLP group members for sharing their research from which I gained numerous inspirations. Particular thanks to Olga Tsubiks for walking me through the data collection process, and April Xu for giving me insights on the financial markets.

Without these people, this thesis would not exist.

Faye Xu

Halifax

August 9th 2012

Chapter 1 INTRODUCTION

1.1 MOTIVATION

Since the introduction of social media, companies are increasingly adopting social media technologies, using Twitter to reach out to customers or YouTube to demonstrate product features. Social media is widely applied as a mean of marketing and public relation by companies. However, online user-generated content (UGC), the real "meat" of social media, can serve businesses far more than marketing. With the help of various data mining tools, researchers are able to view the opinions of large numbers of users publicly. There is enormous amount of contents which indicate the opinions and insights of diverse groups of internet users who possess various types of information and expertise. "The wisdom of crowds", equipped with data mining rules and algorithms can automatically generate collective estimations of future performance on a variety of subjects, such as stock market performance, sports outcomes, election results and box office sales. The opinions and insights are reflected in votes and comments in the forums, sentiments on the tweets, counts of views and ratings on YouTube, etc. In this thesis attempt has been made to study the prediction power of the collective sentiments of micro-blogging websites on the stock market.

In general, a user's activities in social media include authoring content, viewing and networking (Guo, Tan, Chen, Zhang, & Zhao, 2009). In most of the similar research which have been done on stock market prediction, the UGC mainly comes from the internet forums which enable users to bet on and make market predictions about the outcomes of future events(Gu, Konana, Liu, Rajagopalan, & Ghosh, 2006)(Hill & Ready-

Campbell, 2011) (Avery, Chevalier, & Zeckhauser, 2011). In this case, the internet forum is mainly functioning as a venue for users to author content. The participants include active users who constantly engage in stock market prediction, some making more accurate prediction than others; and spamming users who make random guesses. The data from the forums are fairly structured with either positive or negative votes. However, the quality of the votes cannot be traced due to the lack of reliable user profile information in the forums. Moreover, the time-sensitive nature of the financial market requires a medium where information is rapidly spreadable while forum's performance is inferior in this regard. Micro-blogging emerges as an alternative for users to rapidly share their own opinions, and in the meantime actively follow other users' opinions for both information gathering and networking purposes. Many recent studies(Bollen, Mao, & Zeng, 2011)(Pajupuu, Kerge, & Rene, 2011) have been done on Twitter sentiments by using sentiment lexicon to simply count the positive and negative polarity words to correlate the general public sentiment to the stock market index such as Dow Jones Industrial Average (DJIA) or S&P 500. However, no research has been seen by using machine learning algorithms in natural language processing (NLP) to get the public sentiment on individual stocks in order to study its relationship with the stock price change. The use of emoticons, abbreviations, and poor grammar on micro-blogging services like Twitter presents a difficult task for natural language processing. On the other side, the richness of the user information and the availability of real-time data in micro-blogging services impose opportunities for researchers to study on data of better quality.

1.2 RESEARCH PROBLEM FORMULATION

Although each tweet, which is a post made on micro-blogging websites, is limited to 140 characters, there are thousands of millions of tweets generated by hundreds of millions of users every day. With various sentiment detection tools available, it becomes possible to discover knowledge from the relatively unstructured UGC on micro-blogging websites like Twitter. Machine learning enables machines to automate the sentiment detection process and thus take advantage of arbitrage opportunities faster than human counterparts by repeatedly monitoring public sentiments and forecasting price fluctuations in the example of stock market prediction. The data used in the research are tweets collected from stocktwits.com, which is an online financial communication platform for the financial and investing community. At the time of writing, there are more than 150,000 investors on the site, which can be viewed by audiences of 40 million across the financial web and social media platforms. As a sister services to Twitter, StockTwits is composed of a large user base of trading and investing professionals, who can integrate their StockTwits accounts with their Twitter accounts if they choose to. Section 4.2.1 describes StockTwits in more details. In this thesis, various machine learning algorithms are used to detect the public sentiment of the users on StockTwits. Experiments are carried out to study the relationship between the stock market and the public sentiment.

The objectives of the thesis are:

- Determine which natural language processing approaches are the most suitable for sentiment analysis on tweets.
- Study the relationship between the trade volume of the stocks and the activity of users' discussions of the stocks on StockTwits.

- Analyze whether users' collective sentiments of tweets have predictive power on the daily stock market performance.

The ultimate goal of this thesis is of course not to build an ideal model for stock market prediction, but to test whether the feature of social media sentiments contributes to the stock market analysis.

1.3 OUTLINE

In Chapter 2, some trading theories will be introduced as a theoretical base of why the data will be manipulated in such way that makes sense in determining the relationship between the public sentiment and the stock market change. It includes a brief introduction of Efficient Market Hypothesis (EMH), types of stock trading, a list of trading terms which reveal traders' sentiment towards the stock movement, and traders' online networking activities.

In Chapter 3, there will be a discussion of related literature. Exhaustive review of the current literature and other sources will significantly contribute to the further research. The issues and methodologies discovered from the literature will be addressed and referred to throughout the rest of the thesis.

Chapter 4 describes the methodology and the experiment design of the research. The experiments are conducted extensively with a step-by-step approach. Comparative experiments are performed within each individual step and the empirically best performing local solution is carried down to the next step. Section 4.3.5 presents an experimental framework to help illustrate the details involved in the research. In general,

the research architecture mainly consists of two parts: a NLP approach and a statistical analysis approach.

Multiple methods in sentiment detection are discussed in Chapter 5. Three different machine learning algorithms which are used in the analysis in 4.3.3 are introduced. This chapter presents the principle and formulas for the NLP approach and gives algorithmic descriptions in details.

Chapter 6 is composed of a presentation of the results of the analyses and a discussion of the findings. Chapter 7 concludes the thesis and presents some thoughts for future work.

Chapter 2 TRADING THEORY

2.1 EFFICIENT MARKET HYPOTHESIS

Efficient Market Hypothesis (EMH) is an investment theory that states it is impossible to "beat the market" because stock market efficiency causes the current stock prices to always incorporate and reflect all relevant information available in the market. According to the EMH theory, stocks are always traded at their fair value thus making it impossible for investors to either purchase undervalued stocks or sell stocks for inflated prices. For this reason, it should be impossible to outperform the market.

The EMH was developed by Eugene Fama (Fama, 1970). It includes three forms: weak, semi-strong and strong efficiency. Under the weak EMH, future prices cannot be predicted by simply analyzing historical prices. Investors cannot earn returns in the long run by using investment strategies based on historical prices or other historical information. In other words, under weak efficiency technical analysis techniques will not be able to consistently produce excess returns. However, some forms of fundamental analysis may still provide excess returns. The semi-strong form of EMH goes a step further by incorporating all historical and currently public information into the price. The strong form of EMH includes historical, public, and private information, such as insider information, in the share price(Schumaker, Zhang, & Huang, 2011).

According to the EMH stock market prices are largely driven by new information, i. e. news, rather than present and past prices(Bollen, Mao, & Zeng, 2011). Due to the randomness of news, the up and down movement of stocks cannot be predicted with 50% accuracy.

2.2 TYPES OF STOCK TRADING

Based on the duration of stock holding, stock trading can be classified as day trading, short term, medium term, and long term trading. For day trading, both buying and selling of a financial instrument is done on the same day and all the trading is closed before the market closes for the day. Traders who participate in day trading are called active traders or day traders. A trade period of more than one day to a few weeks is considered as short term trade. Medium term trading is with a trade period from a few weeks to a few months. In long term trading, a stock is held for many months to many years.

Fundamental analysis of a business involves analyzing its financial statements and health, its management and competitive advantages, and its competitors and markets. This type of analysis is based on the assumption that markets may misprice a security in the short run but that the fair price will eventually be reached. Profits can be made by purchasing the mispriced security and then waiting for the market to re-price the security to its fair value. Technical analysis, on the other hand, is security analysis for forecasting the direction of prices through the study of past market data, primarily price and volume. It maintains that all information is reflected already in the stock price. Trends and sentiment changes predate and predict trend changes. Investors' emotional responses to price movements lead to recognizable price chart patterns. Technical analysis does not care what the “value” of a stock is. The price predictions of technical analysis are only extrapolations from historical prices. These strategies believe that market timing is critical and opportunities can be found through the averaging of historical price and volume movements and comparing them against current prices(Babu, N.Geethanjali., & Kumari, 2010).

2.3 TRADING TERMS

There are some trading terms which reveal traders' sentiment towards the stock movement. Although it is not possible to list all of them, some of the most common terms are introduced in this section. The appearance of those key terms is proved to be very effective in determining the users' sentiment towards stock movement when training the machine learning algorithms.

Bull/Bear: Bulls are investors who think the stock will rise. Investors who take a bull approach will purchase stocks under the assumption that they can be sold later at a higher price. Bulls are optimistic investors who are presently predicting positively for the market, and are attempting to profit from this upward movement. Bears, in comparison, are pessimistic and believe that a particular stock is headed downward. Bears attempt to profit from a decline in prices. The adjectives bullish and bearish are also frequently used words to describe users' sentiments towards the market.

Long/Short: Long (or Long Position) is the buying of a security such as stock, commodity or currency, with the expectation that the asset will rise in value. For example, an owner of shares in Apple Inc. is said to be "long Apple Inc." or "has a long position in Apple Inc.". Short (or Short Position), which is opposite of "long", is the sale of a borrowed security, commodity or currency with the expectation that the asset will fall in value. For example, an investor who borrows shares of stock from a broker and sells them on the open market is said to have a short position in the stock. The investor must eventually return the borrowed stock by buying it back from the open market. If the stock falls in price, the investor buys it for less than he or she sold it, thus making a profit.

Options – Calls/Puts: An option is common form of a derivative. It's a contract, or a provision of a contract, that gives one party (the option holder) the right, but not the obligation to perform a specified transaction with another party (the option issuer or option writer) according to specified terms.

Call options provide the holder the right (but not the obligation) to purchase an underlying asset at a specified price (the strike price), for a certain period of time. If the stock fails to meet the strike price before the expiration date, the option expires and becomes worthless. Investors buy calls when they think the share price of the underlying security will rise or sell a call if they think it will fall. Selling an option is also referred to as "writing" an option.

Put options give the holder the right to sell an underlying asset at a specified price (the strike price). The seller (or writer) of the put option is obligated to buy the stock at the strike price. Put options can be exercised at any time before the option expires. Investors buy puts if they think the share price of the underlying stock will fall, or sell one if they think it will rise.

Put buyers, those who hold a “long”, are either speculative buyers looking for leverage or "insurance" buyers who want to protect their long positions in a stock for the period of time covered by the option. A put buyer has the right, but not the obligation, to sell the underlying stock at the strike price of the option until the expiration date. Furthermore, if a trader buys a put option, the risk of the trade equals the money paid for the option, or the debit. The profit is equal to the fall in the price of the underlying asset. The profit will result if the underlying security moves lower. The profit is limited because the underlying asset will not fall below zero. Finally, to offset a long put, the trader will sell a put with

the same terms (strike price and expiration) to "close" out the position. On the other hand, if the trader exercises a long put, then he or she is selling, or short, the underlying stock or index at the strike price of the put option.

Put sellers hold a short position expecting the market to move upward. A put seller has the obligation to buy 100 shares (per option) of the underlying stock at the put strike price. In other words, the option seller must be ready to have the stock "put" to him or her. The put seller's risk is the drop in the stock price, which is limited to the stock falling to zero. The profit equals the credit received from the sale of the put. Put sellers often prefer options with little time left until expiration because they want a put to expire worthless. In that way, the seller keeps the entire premium. A short put is offset by purchasing a put with the same strike price and expiration to close out the position (Investopedia).

2.4 TRADERS' ONLINE NETWORKING

Information from breaking news stories can dramatically affect the stock price. However, even when the information contained in financial news articles can have a visible impact on the stock market, sudden price movements can still occur from other sources such as large unexpected trades(Schumaker, Zhang, & Huang, 2011). The ability to predict stock market behavior has always had a certain appeal to researchers. Even though technology has emerged as one of the primary forces shaping trading markets from its inception(Williams, 2011), the difficulty has been the inability to capitalize on the behaviors of human traders(Schumaker, Zhang, & Huang, 2011). Behavioral patterns

have not been fully defined and are constantly changing, thus making accurate predictions quite difficult.

The real-time Web has changed many things for stocks, markets and finance. The social Web and technology advances have evolved to create tools and platforms that will fuel finance and the sharing of ideas. A new concept called social trading, a process through which online financial investors rely on user-generated financial content gathered from various applications as the major information source for making financial trading decisions(Wikipedia, Social trading, 2012), is rapidly emerging. Until recently investors and traders were relying on fundamental and technical analysis to form their investment decisions. Now they can weave into their investment decision process social indicators that are fueled by a transparent real-time trading data feed of all the users in the social trading network. This is now being introduced as social financial analysis. Social trading has also been associated with a variety of online social trading networks, such as Currensee, Zecco and StockTwits (World Finance on Social Trading), which introduce the idea of communal, cooperative and collaborative trading environment.

Based on Lindzon et al.'s research(Lindzon, Pearlman, & Ivanhoff, 2011), successful traders specialize in a favorite setup, which is a combination of factors that need to align in time and space in order to produce a buy or sell signal. Each trader presents a favorite setup in detail as well as the approach of finding new ideas and managing risk that the trader uses successfully on a daily basis. The process of finding new ideas is done either explicitly, by intentionally following the trading activities of one or more selected traders, or implicitly, as one's trading decisions are unintentionally influenced by the trading

activities of other traders. Knowing the market players as well as the market itself is extremely important for traders to ride the trends of stocks.

Normally, electronic Word-of-Mouth (WOM) arises from a possibly unlimited number of unknown participants and the presence of vast amounts of unfiltered information makes the information validity uncertain(Cheung, Luo, Sia, & Chen, Summer 2009). However, in Bakshy and Hofman's paper (Bakshy & Hofman, 2011) about word-of-mouth marketing, they find that the largest cascades tend to be generated by users who have been influential in the past and who have a large number of followers. WOM has long been regarded as an important mechanism by which information can reach large populations, possibly influencing public opinion. When Lazarsfeld and his team researched public communication as early as 1940s, they found out that communication does not directly flows to the mass but is actually interpreted first by opinion leaders and then forwarded to the rest of the people(Lazarsfeld, Berelson, & Gaudet, 1944). Golub and Jackson's recent paper (Golub & Jackson, 2010) studied naïve learning in social networks to see whether all opinions in a large society converge to the truth.

Generally speaking, social networking sites such as StockTwits for traders have emerged as efficient platforms for traders to exchange their ideas and share their insights, thus providing an ideal source for researchers to study human traders' behaviors.

Chapter 3 LITERATURE REVIEW

3.1 TEXT MINING AND THE STOCK MARKET

3.1.1 Formal Channel Text and the Stock Market

There are a variety of prediction techniques used by stock market analysts. Apart from technical analysis in price and volume, there are also non-quantitative factors such as "general regulatory news" which may significantly affect the stock market. Text mining in financial news for stock market prediction was researched by scholars as early as 1990s. Wuthrich et al. built a prediction system that uses data mining techniques and keyword tuple counting and transformation to produce periodically forecasts about stock markets (Wuthrich, et al., 1998).

Babu et al. conducted research on textual analysis of stock market prediction using financial news articles(Babu, N.Geethanjali., & Kumari, 2010). They think that textual data are from two sources: company generated and independently generated sources. Annual and quarterly reports are company generated and they can provide a rich linguistic structure that if properly read can indicate how the company will perform in the future. Independent sources such as analyst recommendations, news outlets, and wire services can provide a more balanced look at the company and have less potential for bias comparing to company generated ones. They believe that discussion boards can also provide independently generated financial news but they can be suspect sources.

Butler and Keselj did a study on financial forecasting using character N-gram analysis and readability scores of annual reports(Butler & Keselj, 2009). Their hypothesis is that vital information is contained in textual content of annual reports for assessing the

performance of the stock over the next year. In their paper, Character N-grams (CNG) distance measure was used to classify annual reports. Readability scores were generated and used together with the previous year's performance to make class predictions using a Support Vector Machine (SVM) method.

Recently, sentiment analysis of financial news articles was carried out by Schumaker et al. to study the impact of objectivity/subjectivity of news article on prediction (Schumaker, Zhang, & Huang, 2011). In general, sentiment analysis is concerned with the analysis of direction-based text, i.e., text containing opinions and emotions. Firstly, they used OpinionFinder to identify the overall tone of the financial news articles to determine whether each article is objective or subjective, and if subjective, whether it is positive or negative. Then they used AZFinText tool to study whether positive/negative subjectivity impact news article prediction. From their findings, they believe that the author's use of subjectivity in the financial news article may have influenced market trading immediately following article release.

3.1.2 Forums, Blogs and the Stock Market

With the introduction of the UGC, many major online portals have independent finance forums which allow users interested in finance to gather and make stock market predictions. In the case of Yahoo! Finance, Gu et al. has done a research on its message board which introduced community stock sentiments to influence investors in their trading decisions (Gu, Konana, Liu, Rajagopalan, & Ghosh, 2006). Their paper studied the predictive power of message board sentiments over future abnormal stock returns. Each post was self-labeled as positive or negative on the prediction of 71 specific

equities. Based on the analysis, a trading strategy that involves buying stocks with low sentiments while selling stocks with high sentiments was implemented.

Unlike Gu et al., Hill and Ready-Campbell introduced an expert stock picker strategy to build investment portfolios (Hill & Ready-Campbell, 2011). They sourced the publicly available votes from a site called CAPS, owned by a financial newsletter publisher Motley Fool. Instead of building a model which includes every vote, they built a model which only includes votes from top stock pickers who have the highest accuracy rate from the past record. They managed to demonstrate that portfolios with the stock picks of a large sample of online users from Motley Fool CAPS can outperform the S&P 500. By applying their approach to identify experts in the crowds of online stock pickers on the site, it is also demonstrated that it is better than letting the entire online crowd vote. There are other ongoing researches that study the stock market prediction power of the "CAPS" website by the Motley Fool company (Avery, Chevalier, & Zeckhauser, 2011).

Choudhury et al.'s paper(Choudhury, Sundaram, John, & Seligmann, 2008) focused on the communication dynamics in the blogosphere where contextual properties of communication can be extracted, such as the number of posts, the number of comments, the length and response time of comments, strength of comments and the different information roles that can be acquired by people (early responders/late trailers, loyals/outliers). Their results were yielding about 78% accuracy in predicting the weekly magnitude of movement and 87% for the weekly direction of movement.

Chen et al. studied sentiment revealed in social media and its effect on the stock market. (Chen, Prabuddha, Hu, & Hwang, 2011)They extracted sentiment by conducting a textual analysis of articles published both through formal channel the Wall Street Journal (WSJ)

and social media site Seeking Alpha. They believe that authors of the articles in Seeking Alpha have a genuine incentive to produce high-quality research reports and increase their network of followers, who later could become their clients and paying subscribers to their financial blogs. Evidence from their research showed that sentiment revealed through Seeking Alpha has a larger and longer-lasting impact on stock returns than views expressed in the WSJ.

3.1.3 Search Volume and the Stock Market

Balthasar's report for Barclays Bank(Balthasar, 2009) studied the questions that whether change in Google search volume has prediction power for stock prices and whether high search volume is correlated to strong momentum. She found that for stock in the NASDAQ 100, the change in search volume cannot be used as a directional indicator for stock prices. However, it is a coincident indicator for stock variance. Although she also did not observe that stocks with high search volume index (SVI) display stronger momentum in a consistent and reliable way, with the realization of that fact that the underlying momentum itself is weak and unstable, she concluded it is possible to discover such correlation in a more persistent momentum strategy.

3.1.4 Micro-blogs and the Stock Market

Bollen et al. investigated whether measurements of collective mood states derived from large-scale Twitter feeds are correlated to the value of Dow Jones Industrial Average over time (Bollen, Mao, & Zeng, 2011). They found an accuracy of 86.7% in predicting the daily up and down changes in the closing values of the DJIA and a reduction of the

Mean Average Percentage Error by more than 6%. Their studies have measured moods in several dimensions (Calm, Alert, Sure, Vital, Kind, and Happy).

Ruiz et al. studied the problem of correlating micro-blogging activity with stock market events(Ruiz & Hristidis, 2012). Features were extracted from micro-blogging platform to measure its overall activity and other properties of its induced interaction graph, for instance, the number of connected components, statistics on the degree of distribution and other graph-based properties. They found that the most correlated features are the number of connected components and the number of nodes of the interactive graph. The correlation is stronger with the traded volume than with the price of the stock. They also conducted a simulation by taking into account the relatively small correlations between price and micro-blogging features and found that it drove a stock trading strategy that outperforms other baseline strategies.

3.2 Literature in Natural Language Processing for Sentiment Detection

Sentiment analysis, the process of automatically detecting if a text segment contains emotional or opinionated content and determining its polarity, is a field of research that has received significant attention in recent years, both in academia and in industry. Pang et al. (Pang, Lee, & Vaithyanathan, 2002) were among the first to explore the sentiment analysis of reviews focusing on machine learning approaches. They experimented with three different algorithms: Support Vector Machine (SVM), Naive Bayes (NB) and Maximum Entropy (ME) classifiers, using a variety of features, such as unigrams and bigrams, part-of-speech (POS) tags and binary and term frequency feature weights. Their best accuracy attained in a dataset consisting of movie reviews used a SVM classifier

with binary features, although all three classifiers gave very similar performance. Later, the same authors presented an approach based on detecting and removing the objective parts of documents (Pang & Lee, 2004). The results showed an improvement over the baseline of using the whole text using a NB classifier but only a slight increase compared to using a SVM classifier on the entire document.

Satapathy and Bhagwani's work(Satapathy & Bhagwani, 2012) aims to infer emotions from sentences using a lexicon-based approach. Unigrams are used as major emotion indicators. In order to capture lower order dependencies, which unigrams fail to capture, bigrams and trigrams are used. Their work focuses on two tasks: a Coarse-Grained classification of the sentences based on their polarity and a Fine-Grained classification which aims at inferring the emotion conveyed by the sentence, based on a pre-defined list of emotions (joy, anger, fear, disgust, and guilt). Bollen et al. took similar approach in this regard by measuring mood in terms of 6 dimensions and they found that the accuracy of DJIA predictions can be significantly improved by the inclusion of specific public mood dimensions but not others (Bollen, Mao, & Zeng, 2011).

Many approaches to automatic sentiment analysis begin with a large lexicon of words marked with their prior polarity(Chen, Prabuddha, Hu, & Hwang, 2011)(Bollen, Mao, & Zeng, 2011)(Baccianella, Esuli, & Sebastiani, 2010). These kinds of approaches have been used as the de facto standard of much research especially financial article research primarily because of its simple nature and ease of use(Babu, N.Geethanjali., & Kumari, 2010). However, Wilson et al. think that the contextual polarity of the phrase in which a particular instance of a word appears may be quite different from the word's prior polarity(Wilson, Wiebe, & Hoffmann, 2009). Positive words might be used in phrases

expressing negative sentiments, or vice versa. Also, quite often words that are positive or negative out of context are neutral in context, meaning they are not even being used to express a sentiment. Their work aimed to automatically distinguish between prior and contextual polarity, with a focus on understanding which features are important for this task. Because an important aspect of the problem is identifying when polar terms are being used in neutral contexts, features for distinguishing between neutral and polar instances are evaluated, as well as features for distinguishing between positive and negative contextual polarity. They evaluated how the presence of neutral instances affects the performance of features for distinguishing between positive and negative polarity. Their experiments showed that the presence of neutral instances greatly degrades the performance of these features, and that perhaps the best way to improve performance across all polarity classes is to improve the system's ability to identify when an instance is neutral.

Chapter 4 METHODOLOGY AND EXPERIMENT DESIGN

Ruiz's et al.'s research(Ruiz & Hristidis, 2012), Bollen et al's research(Bollen, Mao, & Zeng, 2011), and Hill and Ready-Campbell's research(Hill & Ready-Campbell, 2011) are of particular interest to the formation of the ideas in the thesis. Hill and Ready-Campbell's applied genetic algorithm approach and identified the online participants who have the highest accuracy rate from the past record as experts, whose opinions are exclusively taken into account to build investment portfolios. In the context of a networking oriented social media, the experts, or the opinion leaders in the case of Twitter are made explicit to the public by their numbers of followers and times of retweets of their posts. Bollen et al's research, on the other hand, treated each tweet equally and calculated the ratio of positive vs. negative messages for the tweets posted on the same day. Ruiz et al.'s research took a bold approach to study the problem of correlating micro-blogging activities with the performance of individual stocks. Volume of related tweets and other features which measure properties of an induced interaction graph were taken into account to study the correlation. In this thesis, an attempt is made to quantify the influence of the tweets by putting into the equation the measurements of experts' followers to reflect their public influence. Meanwhile, it is the individual stocks that are studied in the thesis rather than the stock market index.

4.1 RATIONALE OF MODELING

In this thesis, the models are trained from a corpus of hand-labeled data instead of using sentiment lexicons, such as the SentiWordNet. The SentiWordNet is a great resource of determining twitter sentiment and has been used in multiple research papers. Previous

research as typically used SentiWordNet for the following three tasks, all of which are about tagging a given text according to expressed opinion(Baccianella, Esuli, & Sebastiani, 2010).

1. Determining text SO-polarity, as in deciding whether a given text has a factual nature (i.e. describes a given situation or event, without expressing a positive or a negative opinion on it) or expresses an opinion on its subject matter. This amounts to performing binary text categorization under categories Subjective and Objective (Pang & Lee, 2004).
2. Determining text PN-polarity, as in deciding if a given Subjective text expresses a Positive or a Negative opinion on its subject matter (Pang & Lee, 2004).
3. Determining the strength of text PN-polarity, as in deciding e.g. whether the Positive opinion expressed by a text on its subject matter is Weakly Positive, Mildly Positive, or Strongly Positive(Pang & Lee, 2005).

There are several reasons why existing lexicons are not used in this thesis. First of all, although subjectivity and objectivity are taken into account in determining the tweet sentiment, in many cases objective tweets under normal circumstances should be treated as polarized tweets in this thesis. For example, the tweet *"Added to positions like $AA and $BAC...."*, which will most likely be marked as objective and thus be excluded from the polarity analysis, actually implies that a user expects the stock price to rise. Secondly, the vast majority of work in sentiment analysis mainly focuses on the domains of movie reviews, product reviews and blogs, in which cases SentiWordNet represents a suitable lexicon; however, recent financial research shows that the word lists developed for other disciplines misclassify common words in financial texts (Loughran & McDonald, 2011).

For example, for the tweet *"Short $AAPL @557.50"* , if regular lexicons are used, the sentiment will probably be marked as objective or neutral, while in finance the word *short* is a clear sign indicating that the user expects the $AAPL stock to fall. Thirdly, it is hard to tell the strength of the text polarity from the short texts in tweets, and it is believed that it will be sufficient to mark simple positivity and negativity.

4.2 INSTRUMENTS

4.2.1 StockTwits

StockTwits, as previously mentioned, is used as the source of the tweets to be analyzed. StockTwits content is focused solely on investing, and its technology and staff work to filter out unrelated messages and spam, ensuring the remaining content contains only the most relevant discussions specifically about stocks and markets. It is believed that the users on StockTwits have the genuine incentive to produce high-quality tweets in order to increase their network of followers. StockTwits created the $TICKER tag to enable and organize "streams" of information around stocks and markets across the web and social media such as Twitter, given that the users connected their Twitter account with the StockTwits account. Each tweet is a stream of text with a limit of 140 characters. Each user has a certain number of followers and is following a certain number of users, from whom he/she can get his/her insights at real time by reading their tweets.

Through the initial collection and analysis of tweets of 64 stocks, the list of stocks to be included in the analysis has been narrowed down to 16 stocks, which are the most discussed ones on StockTwits. The list of the stocks included in the analysis can be found in Table 1.

Table 1: List of Stocks

$TICKER	Company Name	Stock Market
$AAPL	Apple Inc.	NASDAQ
$AMZN	Amazon.com, Inc.	NASDAQ
$BAC	Bank of America Corp	NYSE
$BIDU	Baidu.com, Inc.	NASDAQ
$C	Citigroup Inc.	NYSE
$CMG	Chipotle Mexican Grill, Inc.	NYSE
$FSLR	First Solar, Inc.	NASDAQ
$GOOG	Google Inc	NASDAQ
$GS	Goldman Sachs Group, Inc.	NYSE
$IBM	International Business Machines Corp.	NYSE
$JPM	JPMorgan Chase & Co.	NYSE
$MSFT	Microsoft Corporation	NASDAQ
$NFLX	Netflix, Inc.	NASDAQ
$PCLN	Priceline.com Inc	NASDAQ
$RIMM	Research In Motion Limited (USA)	NASDAQ
$YHOO	Yahoo! Inc.	NASDAQ

Tweets of those stocks are collected over a period of two and a half months, along with the information of the username, date and time of publishing, source of the message, and other information which reflects the publisher's profile. A complete list of the initially collected attributes can be found in Table 2. Other attributes will be derived from this list of attributes.

Table 2: List of Attributes for Collection

Entity	Attributes for Collection
Tweets	Username, content of the tweets, date of publishing, time of publishing, source of the message
Users	Username, date of joining, total number of tweets by the user, number of followers, level of experience, approach of trade, holding period of trade

4.2.2 TagHelper

Stock tweets' sentiment detection is mainly achieved by using TagHelper tool (Rosé, et al., 2007), which is an application that makes use of the functionality provided by the Weka toolkit -- a machine learning software written in Java and is licensed under the GNU General Public License. TagHelper's basic classification functionality is available to researchers to use in their own work. It is publicly available and is designed to work with pre-segmented English, German, or Chinese texts, although additional licensing is required for enabling the Chinese functionality. It has been observed that the texts from StockTwits are all in English, so TagHelper is an ideal tool for building text classification models for tweets from StockTwits. To set up the data for analysis, a total number of 2380 tweets are hand-labeled as the training and testing data for the classification models. The hand-labeling process will be described in 4.3.2. Once the training and testing data is prepared and loaded into TagHelper, a number of customizations are performed to maximize the classification performance. Models are trained using all of the hand-labeled tweets from the dataset that will then be used to apply labels to any unlabeled tweets in that set. The training data, which is the hand-labeled tweets, will also be cross validated to estimate the level of performance that can be expected of the trained model on

unlabeled data. To avoid over-fitting, the model is tested on a set of testing data which is also hand-labeled.

In order to use the customizations available in TagHelper tools purposefully it is important to understand that machine learning algorithms induce rules based on patterns found in structured data representations. There are various machine learning algorithm options available through the Weka toolkit, upon which TagHelper is built. Three options are used in this thesis, including Naïve Bayes (NB), which is a probabilistic model, SMO, which is Weka's implementation of Support Vector Machines (SVM), and J48, which is one of Weka's implementations of a Decision Tree (DT) learner. Chapter 5 presents the details of how machine learning algorithms are applied to the NLP tasks.

4.2.3 Statistical and Database Tools for Analysis

MS ACCESS: MS ACCESS is used to aggregate the daily sentiments on the 16 stocks. Multiple approaches are taken to calculate the collective sentiments for each day. The aggregation process is described in detailed in 4.3.4.

SPSS: SPSS is used as a basic statistical tool to study the relationship between the tweets and the stock market. Initial analysis is carried out by using SPSS before further analysis is conducted. Each pair of series of the 16 stocks is studied individually. SPSS is also used as the tool to carry out Binary Logistic Regression.

Eviews: Eviews is a statistical package used mainly for time-series oriented analysis. Granger Causality (GC) test (Granger, 1969) is carried out by Eviews to study whether relationship exists between the stock price change and users' collective sentiments of tweets. It is not to test the actual causation, but whether one time series has predictive information about the other. Section 6.2.2 contains a brief introduction of the GC test.

4.3 DATA ANALYSIS

4.3.1 Data Acquisition and Pre-processing

Stock tweets: The tweets from StockTwits are of good quality in terms of their relevance to the topic. In Ruiz et al's research(Ruiz & Hristidis, 2012), it was difficult to extract relevant stock tweets of companies like Yahoo ($YHOO) and Apple ($AAPL) from Twitter, due to the fact that those names are frequently mentioned for purposes other than stock price discussions. For instance, the short name for Yahoo is used in many tweets that are related with the news service provided by the same company (Yahoo! News). In the second case, Apple is a common noun and is also used widely for spamming purposes (e.g. "Win a free iPad" scams). Although there are fewer tweets available on StockTwits, the relatively good quality of the tweets makes it possible to take into account every tweet. Tweets of the 16 stocks, which are randomly selected and are actively discussed by StockTwits users every day, are collected for the period between March 13th, 2012 and May 25th, 2012. The period chosen did not have unusual market conditions and was a good test bed for the evaluation.

The following steps are taken to pre-process the stock tweets:

- Remove tweets on weekends and public holidays.
- Remove duplicated tweets by the same user.
- Replace @ sign in the tweets with text "atreplace".
- Replace $TICKER for each stock in the tweets with text "stocksignreplace".
- Replace links in the tweets with text "linkreplace".
- Convert all the tweets text to be lowercase.

There are approximately 100,000 tweets for the 16 stocks included in the final analysis. In some research(Ruiz & Hristidis, 2012), the effect of trend was considered because the number of tweets is increasing due to Twitter's increasing user base during the observation period. In this analysis, the trend effect is not taken into account since the period is less than three months. If the data is collected over a longer period, the feature values can be normalized with a time-dependent normalization factor that considers trends.

Stock market data: the stock data is obtained from Google Finance for the 16 stocks for the period between March 13th, 2012 and May 31st, 2012. More days are included in this set of data since the days are lagged to test the prediction power of the stock tweets. For each stock, the daily open and close price and daily traded volume are also recorded over the same period of time. The price series are then transformed into its daily relative change, i.e., if the series of price is P_i , the daily relative change will be $\frac{P_i - P_{i-1}}{P_{i-1}}$.

4.3.2 Tweet Sentiment Hand-labeling

A total number of 2380 tweets on all of the 16 stocks are hand-labeled as the training and testing data. The tweets are labeled as positive (1), negative (-1) or neutral (0). The hand-labeled data are then randomized to generate training dataset and testing dataset, which are of the amount of 2000 and 380, respectively.

For some tweets, it is difficult to label its polarity even by human efforts. In the case of vagueness, the tweets are labeled as neutral. Table 3 presents some typical tweets and their labels as well as the arguments of why they are labeled so.

Table 3: Some Typical Tweets and Their Labels

Tweets	Argument	Label
The one thing I see here on $CMG is that it has touched and rallied from 401 four times since 4/24, each time making a lower high.	One can argue that it can be negative sentiment towards $CMG over the long term. However, one cannot be certain thus it is marked as neutral.	Neutral
Thinking banks will rebound soon.. $JPM	The statement is short and the positive sentiment towards $JPM is clear.	Positive
$AMZN is the paint dry yet?	"Paint dry" is a phrase which describes the process is slow and boring. The question mark reveals that the user is uncertain about it.	Neutral
$AMZN looks good to short fundamentally from tablet risk and subsidized shipping #amazon #kindle $UPS http://t.co/l2KJ3m6K	Although the word *good* usually represents positive sentiment, the word *short* is the key polarity word in this case and it is a clear sign of negative sentiment towards $AMZN.	Negative
$C 6,3% $BAC 6,26% $RIO 3,95% $DANG 7,22% $YOKU - 1,41% $GOOG 2,09% $BIDU 1,73% $SOHU 8.08% $WFM 2,59% $SBUX 2,41%	In this tweet, there are lots of numbers which are the daily percentages of changes of several stocks. However, it is merely a description of the situation rather than positive or negative sentiment.	Neutral

Some general rules are applied when labeling the data.

- If the tweet contains external links of long articles or numerical charts about the stocks, it is generally marked as neutral. The content of the article and the information revealed by the chart are not taken into account.
- Positive or negative labels are only given when the sentiment can be explicitly speculated from the tweet.
- Tweets with question marks are generally marked as neutral.
- Simple summarizations of the stock performance by the end of the day are not taken into consideration.
- If the user reports a loss in a subjective way instead of reporting numbers, it is fair to assume that the user has a negative feeling towards the stock; and vice versa.

4.3.3 Sentiment Detection

Models are trained and cross-validated by using the 2000 tweets, and then tested on the 380 tweets. Data is prepared before applying machine learning algorithms to train the models. Tokenization, a process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens, is applied to the dataset. The list of tokens becomes input for further processing. In general, the standard attribute space is set up with one binary attribute per unique word (unigram) or per unique pair of words (bigrams) in the corpus such that if that word or pair of words occurs in a text, its corresponding attribute will get a value of 1, otherwise it will get a value of 0(Rosé, et al., 2007). Other features are added in order to bias the algorithms to learn rules that are more valid or conceptually more closely related to the definitions of the target categories.

The following is a description of the features used for training the models.

- ***Punctuation:*** Punctuation is treated as a feature in determining whether the tweet is neutral or polarized, but not used as a feature to determine the positivity and negativity. For example, punctuation such as the question mark is a useful indicator of the uncertainty of the statement. Furthermore, the inclusion of a comma might mark that a contribution is relatively more elaborated than one without a comma.
- ***Line length:*** Line length is used as an attribute because it is believed that length of contribution can sometimes serve as a proxy for depth or level of detail. In this case, lengthy contributions in tweet data contain elaborated explanations, which are important to detect in determining whether it is a neutral statement or a polarized declaration.
- ***Unigrams and bigrams:*** A unigram is a single word, and a bigram is a pair of words that appear next to one another. Unigrams are the most typical type of text feature. Bigrams may carry more information, such as lower order dependencies, which unigrams fail to capture(Satapathy & Bhagwani, 2012). For example, bigrams capture the contradictory meaning of the word *long* results between the phrases "long puts" and "long $AAPL". To keep the size of the feature space down, which aids in effective rule learning, rarely occurring features are removed as a simple way of stripping off features that are not likely to contribute to useful generalizations. The threshold of feature removal is five features in determining whether the tweet is neutral or polarized, and two features in determining the positivity and negativity due to its smaller sample size. Meanwhile, stop words are removed because they are irrelevant in determining the tweet sentiment.

However stemming, which is a technique for removing inflection from words in order to allow some forms of generalization across lexical items, is not used because of the sensitivity of some financial words' meanings due to stemming.

Since the grammar in tweets is generally poor, part-of-speech (POS) tagging is not used in the models.

4.3.4 Multiple Approaches in Determining Collective Sentiments

By applying machine learning in NLP, each tweet is labeled with positive (1), negative (-1) or neutral (0). Multiple approaches are taken to determine the collective sentiments of each stock for each day. The following parameters, which are count of tweets, number of followers, and time of publishing, are adjusted in multiple ways to calculate the collective sentiments.

Count of tweets: If a user posted several tweets about a certain stock on the same day, those tweets are aggregated to generate the user's unified sentiment, which is positive (1), negative (-1) or neutral (0). This can avoid counting the tweets with the same sentiment from the same user for multiple times.

Number of followers: In one case, the sentiments of multiple users of the same stock for the same day are added up to generate the collective sentiment; in the other case, different users' unified sentiments are assigned with different weights by taking into account their numbers of followers.

Time of publishing: In light of Schumaker et al.'s research, in which the collection period of financial articles was restricted to be between the hours of 10:30am and 3:40pm because they felt it important to reduce the impact of overnight news on stock prices(Schumaker, Zhang, & Huang, 2011), in this thesis similar considerations are taken

by aggregating the sentiments of different periods. In one case, the sentiments of the tweets posted during 12:00am and 11:59pm on the same day are aggregated, while in the other case, the sentiments of tweets posted during the period of 4:00pm (the marketing closing time) and the next day 9:30am (the market opening time) are aggregated. It is believed that the sentiments during the open time of the market may be influenced by the real-time market fluctuations, thus create noises in the sentiments' prediction power for the future. The sentiments during the closing time, however, are most likely based on the users' logical and intuitive analysis of data and factual information.

4.3.5 Schema

Figure 1 represents a brief research schema of sentiment detection. As previously mentioned there is a total number of 2380 tweets on all of the 16 stocks as the training and testing data. Each tweet is equally treated by replacing the $TICKER for each stock with the identical text. And the randomization process ensures that each tweet has equal chance to appear in the training data or the testing data. Out of the 2000 tweets of training data, 1028 of the tweets are labeled either positive or negative and 972 of them are labeled neutral; while out of the 380 testing data, the numbers are 181 and 199, respectively. The sentiment detection process is composed of two stages. In the first stage, tweets are classified to be either neutral or polarized. The polarized tweets are then carried to the second stage to classify whether they are positive or negative.

Figure 1: Research Schema of Sentiment Detection

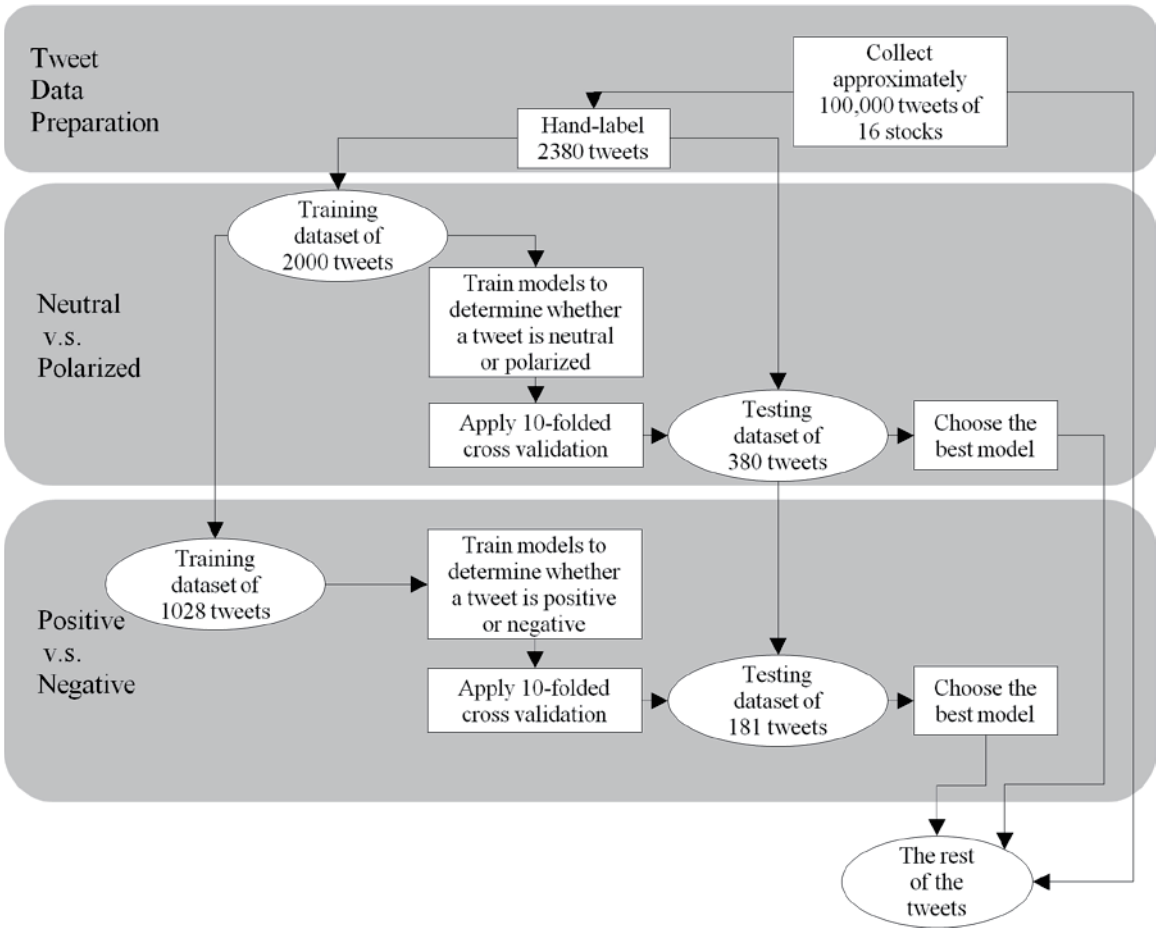
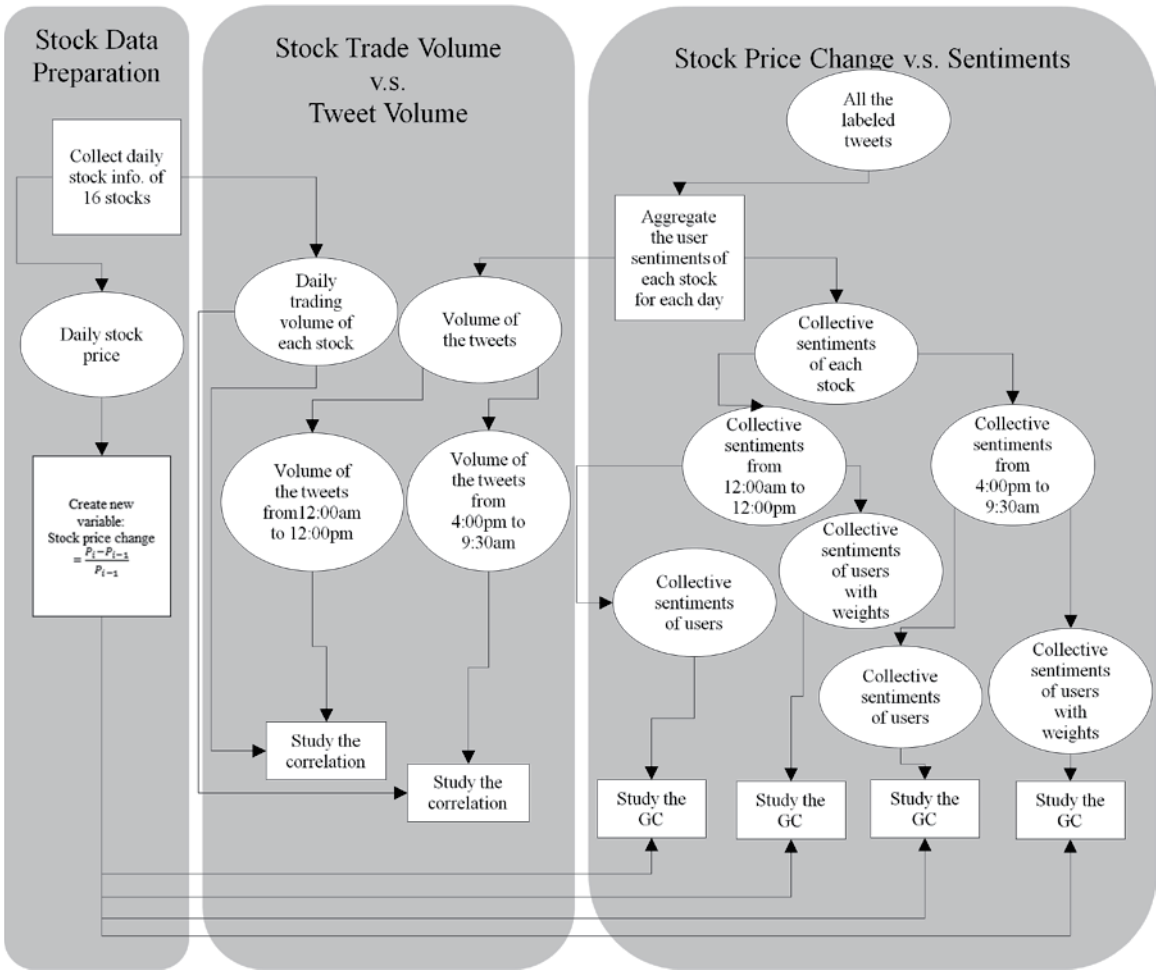


Figure 2 is a description of the analytical process after the data are labeled. It mainly includes two types of analyses, which are the relationship between the stock trade volume and the tweet volume, and the relationship between stock price change and collective sentiments. As mentioned in the previous section, there are multiple approaches in determining collective sentiments. Each collective sentiment generated by different approaches is studied individually with the stock market price change. In Figure 2, *GC* stands for Granger Causality.

Figure 2: Research Schema of Data Analysis



Chapter 5 MULTIPLE METHODS IN SENTIMENT DETECTION

Three machine learning classifiers, which are Naïve Bayes, Decision Tree (J48 in Weka), and Support Vector Machine (SMO in Weka), are applied to the sentiment detection process, which is composed of two stages: Neutral v.s. Polarized Detection and Positive v.s. Negative Detection. Apart from unigrams and bigrams, punctuation and line length are also used as features in the classifiers.

Figure 3 shows the top 50 features of Neutral v.s. Polarized detection according to the Chi-squared attribute evaluation. Punctuations have shown to be very strong indicators of whether a tweet is neutral or polarized, with question mark being the leading feature of all. The feature of whether an external link is included in the tweet also appears to be a strong indicator. Other unigrams such as "short", "long", "calls" and "puts", and bigrams such as "short stock", "stock calls", "BOL sold", and "BOL short" also appear to be strong indicators as expected. Line length is a good indicator as well.

Figure 3: Top 50 features of Neutral v.s. Polarized Detection

QUESTION_MARK, linkreplace, short, long, fb, COLON, calls, DASH, puts, BOL_stocksignreplace, google, short_stocksignreplace, got, line_length, facebook, fb_stocksignreplace, sold, stocksignreplace_calls, msft, holding, bear, laggard, strong, atreplace, aapl, weekly, tlt, recap, nice, good, EXCLAMATION_MARK, PERIOD, PERCENT_SIGN, video, red, billion, new_post, nasdaq, es, links, BOL_sold, bull, BOL_short, gmcr, BOL_long, DOLLAR, play, banks, position, results, …

Figure 4 shows the top 50 features of Positive v.s. Negative detection according to the same Chi-squared attribute evaluation. Punctuations are excluded from the features since

they actually decrease the accuracy rate during the experiment. As expected, unigrams and bigrams such as “short”, “puts”, “short stock”, “long”, “bear”, “red” are on the top of the list. Line length does not appear to be a significant contributor any more. It is worth noting that many stock names also appear to be good predictors. They are not the stocks which are being predicted but are stock names which appear together with the stocks of prediction. The reason might be very complicated, for example, “aapl” might be mentioned frequently with all other stocks for comparison purposes since it has always been a bullish stock for the past few years; or it might be mentioned together with other stocks to be on the users’ long (or short) list.

Figure 4: Top 50 features of Positive v.s. Negative Detection

short, puts, amzn, short_stocksignreplace, gs, long, ma, nice, jpm, xlf, june, bear, red, BOL_short, c, lower, long_stocksignreplace, shorting, higher, wfc, 8, 590, small, BOL_long, bac, buying, run, weekend, laggard, nflx, aapl, fas, bank, usually, boys, big_boys, people, sold, crisis, stocksignreplace_june, BOL_rt, bull, resistance, job, lvs, banks, pcln, goog, risk, sales, …

5.1 NAÏVE BAYES CLASSIFIER

The NB classifier will consider each of these attributes separately when classifying a new instance and it works under the assumption that one attribute works independently of the other attributes contained by the sample. NB is a simple classification method based on Bayes rule, which is represented by the following formula:

$$P(c|t) = \frac{P(t|c)P(c)}{P(t)}$$

where *t* is the tweet and *c* is the class of the tweet.

$$\mathrm{C_{MAP}} = \underset{\mathrm{c}\epsilon\mathrm{C}}{\operatorname{argmax}}\, \mathrm{P}(x_1, x_2, \ldots, x_\mathrm{n}\,|\mathrm{c})\mathrm{P}(\mathrm{c})$$

In the formula above, x_1, x_2, x_n are the features extracted from the training dataset. C_{MAP} represents the most likely class (MAP is "maximum a posteriori"). The NB classifier makes use of all the attributes contained in the data, and analyses them individually as though they are equally important and independent of each other.

5.2 DECISION TREE CLASSIFIER

J48 is an open source Java implementation of the C4.5 algorithm in the Weka data mining tool. C4.5 builds decision trees from a set of training data using the concept of information entropy. The training data is a set T = $t_1, t_2, \ldots\ t_n$ of already classified tweets. Each sample t_1= $x_1, x_2, \ldots x_n$ is a vector where $x_1, x_2, \ldots x_n$ present attributes or features of the tweet. The training data is augmented with components C = $c_1, \mathrm{c}_2, \ldots c_n$ where $c_1, \mathrm{c}_2, \ldots c_n$ represent the class to which each tweet belongs.

At each node of the tree, C4.5 chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. Its criterion is the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm recursively classifies until each leaf is pure, meaning that the data has been categorized as close to perfectly as possible. In other words, among the possible values of a feature, if there is any value for which there is no ambiguity, that is, for which the data instances falling within its category have the same

value for the target variable, then that branch is terminated and assigned with the obtained target value(Aggarwal & Aggarwal, 2011). This process ensures maximum accuracy on the training data, but it may create excessive rules that only describe particular idiosyncrasies of that data. When tested on new data, the rules may be less effective, which is proved to be the case as described section 6.1.

Figure 5 and Figure 6 are small fractions of the decision trees constructed for Neutral v.s. Polarized detection and Positive v.s. Negative detection with 146 leaves and 49 leaves, respectively.

Figure 5: Decision Tree of Neutral v.s. Polarized Detection

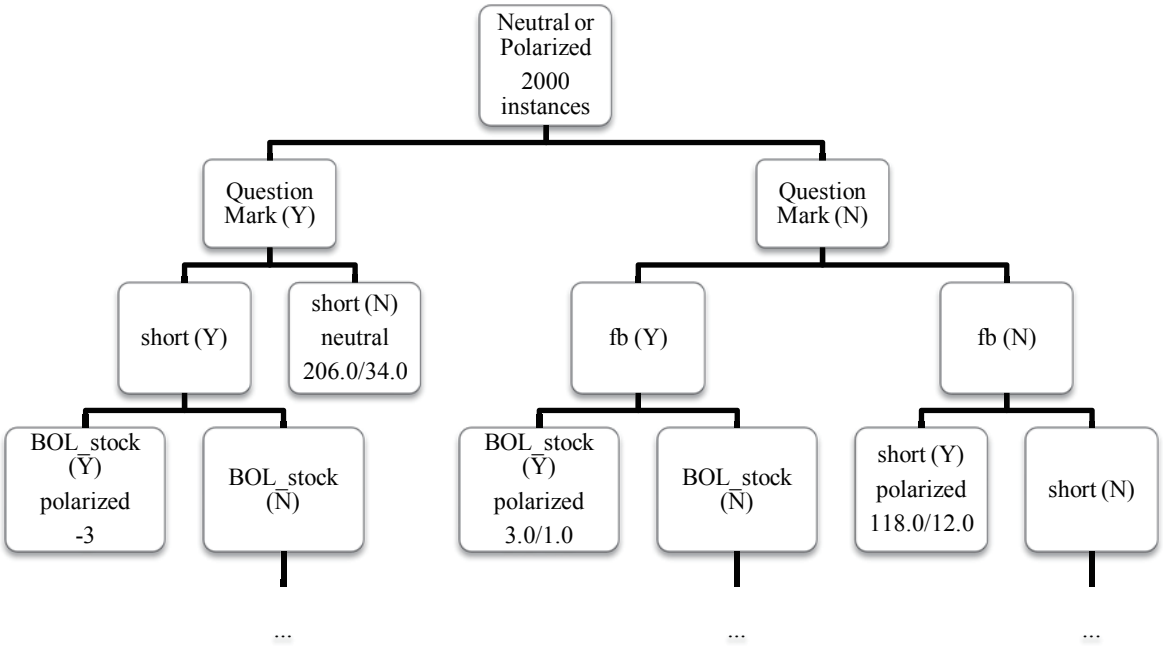
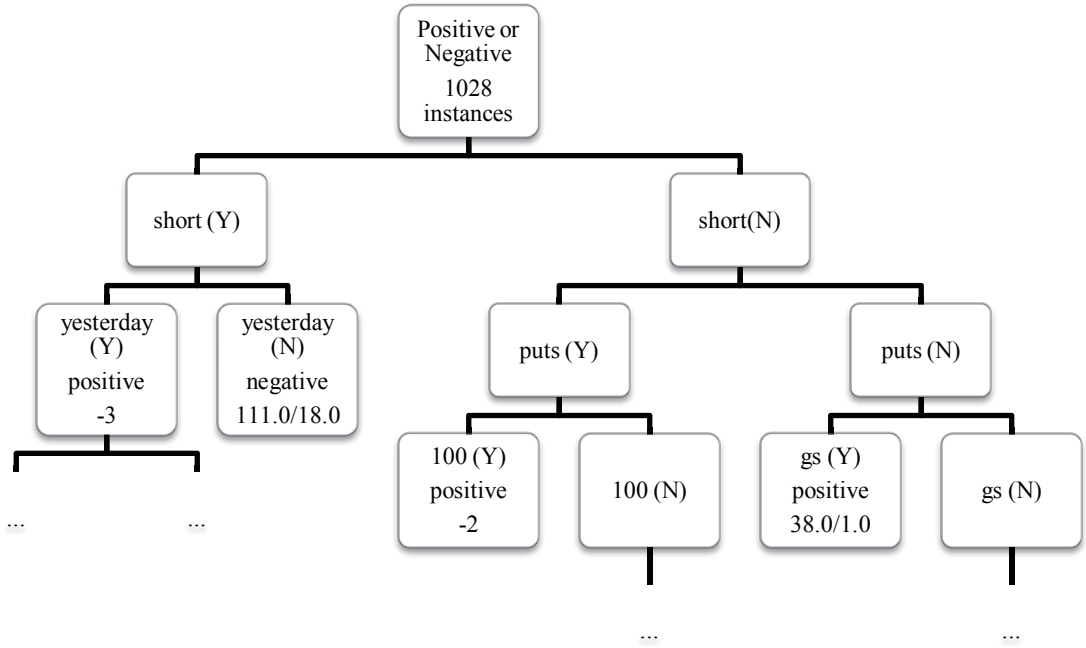


Figure 6: Decision Tree Positive v.s. Negative Detection



5.3 SUPPORT VECTOR MACHINE CLASSIFIER

An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. The advantage of SVM is that they can make use of certain kernels in order to transform the problem, such that linear classification techniques can be applied to non-linear data. Applying the kernel equations arranges the data instances in such a way within the multi-dimensional space, that there is a hyper-plane that separates data instances of one kind from those of another.

Sequential minimal optimization (SMO), which is used in this thesis, is an algorithm for efficiently solving the optimization problem which arises during the training of SVM.

This implementation globally replaces all missing values and transforms nominal attributes into binary ones. It also normalizes all attributes by default.

5.4 Reasons of Choosing the Classifiers

Naïve Bayes classifier is fast, accurate, simple, and easy to implement, thus chosen to be one of the classifiers in this case. It is based on a simplistic assumption in real life and is only valid to multiply probabilities when the events are independent. Despite its naïve nature, NB classifier actually works well on actual datasets.

Decision trees are a classic way to represent information from a machine learning algorithm, and offer a fast and powerful way to express structures in data. Decision trees require little data preparation and are easy to understand and interpret. It uses a white box model, which contrasts to a black box model such as artificial neutral network, meaning that the explanation for the condition is easily explained by boolean logic.

SVM classifiers have been considered state-of-the-art for sentiment classification by many research papers (Pang, Lee, & Vaithyanathan, 2002)(Schumaker, Zhang, & Huang, 2011)(Zhai, Cohen, & Atreya, 2011). Practical real-world modelers have found that SVMs have performed well when other classifiers did poorly. SVM classifiers have been widely used in text classification tasks with unbalanced training. It is also chosen to be used in this thesis.

Chapter 6 RESULT ANALYSIS AND DISCUSSION

The research architecture of the thesis mainly consists of two modules: a NLP approach to label the sentiment of each tweet and an analytical component to study the prediction power of the stock tweets' sentiment on the stock market. The results will be presented accordingly in this chapter.

6.1 RESULT OF SENTIMENT DETECTION

The sentiment detection process is a two-stage process. Firstly, tweets are classified to be either neutral or polarized. Then the polarized tweets are carried to the second stage to classify whether they are positive or negative. The overall accuracy rates of the best classifiers at the two stages are 71.84% and 74.3%, while the numbers for ZeroR baseline models are merely 51.4% and 57.98%, respectively, These are satisfying results, as the process of determining the sentiment of a tweet is vague even for human as described in 4.3.2.and people only agree on sentiment 80% of the time(Grimes, 2010).

6.1.1 Neutral v.s. Polarized Detection

For the training data, Support Vector Machine appears to be the model with the highest overall accuracy. At this stage, it is important to raise the discussion about the F measure since it is both the precision and recall that should be looked into in order to carry the best set of polarized labeled data into the next stage. If the precision of predicting polarized tweets is more valued than the recall, many misclassified polarized tweets which are supposed to be carried to the next stage will be dropped; if the recall is more valued, many misclassified neutral tweets will be mistakenly carried to the next stage. It is arguable that whether the precision or recall is more important at this stage, so it is fair

to assign equal weights to them in the F measure. In Table 4, it becomes evident that SVM has both the highest overall accuracy as well as the highest F measure in the training dataset. For the testing results in Table 5, it seems that Naïve Bayes has an F measure slightly higher than SVM's although SVM still has the highest overall accuracy. Since the difference between the F measures is very small, SVM is chosen to the classifier to apply to the rest of the dataset.

Table 4: Training Results of Neutral v.s. Polarized Detection

Machine Learning Algorithm	Overall Accuracy	Precision of Polarized tweets	Recall of Polarized tweets	F measure	Classification Matrix
Naïve Bayes	67.70%	65.97%	76.75%	0.7095	a b <-- classified as 565 407 | a = neutral 239 789 | b = polarized
Decision Tree	67.25%	67.22%	70.82%	0.6897	a b <-- classified as 617 355 | a = neutral 300 728 | b = polarized
Support Vector Machine	70.50%	70.98%	72.08%	0.7153	a b <-- classified as 669 303 | a = neutral 287 741 | b = polarized

Table 5: Testing Results of Neutral v.s. Polarized Detection

Machine Learning Algorithm	Overall Accuracy	Precision of Polarized tweets	Recall of Polarized tweets	F measure	Classification Matrix
Naïve Bayes	68.95%	64.00%	79.56%	0.7094	a b <-- classified as 118 81 | a = neutral 37 144 | b = polarized
Decision Tree	66.05%	62.04%	74.03%	0.6751	a b <-- classified as 117 82 | a = neutral 47 134 | b = polarized
Support Vector Machine	71.84%	70.56%	70.17%	0.7036	a b <-- classified as 146 53 | a = neutral 54 127 | b = polarized

6.1.2 Positive v.s. Negative Detection

For the polarity detection, as shown in Table 6, SVM again has the highest overall accuracy and highest F measure in the training dataset. In the testing dataset, as shown in Table 7, SVM's overall accuracy is not as high as the NB model's; however, it still has the highest F measure. As a result, it is still chosen to be the best classifier. The two SVMs models are then applied to the rest of the dataset in sequence.

Table 6: Training Results of Positive v.s. Negative Detection

Machine Learning Algorithm	Overall Accuracy	Precision of Positive tweets	Recall of Positive tweets	F measure	Classification Matrix
Naïve Bayes	71.01%	75.87%	73.32%	0.7457	a b <-- classified as 293 139 | a = negative 159 437 | b = positive
Decision Tree	68.00%	68.52%	82.89%	0.7502	a b <-- classified as 205 227 | a = negative 102 494 | b = positive
Support Vector Machine	75.68%	77.12%	82.55%	0.7974	a b <-- classified as 286 146 | a = negative 104 492 | b = positive

Table 7: Testing Results of Positive v.s. Negative Detection

Machine Learning Algorithm	Overall Accuracy	Precision of Positive tweets	Recall of Positive tweets	F measure	Classification Matrix
Naïve Bayes	74.59%	84.21%	72.07%	0.7767	a b <-- classified as 55 15 | a = negative 31 80 | b = positive
Decision Tree	68.51%	69.57%	86.49%	0.7711	a b <-- classified as 28 42 | a = negative 15 96 | b = positive
Support Vector Machine	74.03%	76.23%	83.78%	0.7983	a b <-- classified as 41 29 | a = negative 18 93 | b = positive

6.2 RESULTS OF THE ANALYSES

Different types of correlation analyses are carried out on each of the 16 stocks and the procedures are as shown in 4.3.5. The correlation between the volume of the tweets by unique users and the daily trading volume is studied in two different approaches, which are varied due to the different definitions of time periods as described in 4.3.4.

The other analyses study the relationship between the collective sentiments of users and the daily stock price change, and the variances between the models lie in the different definitions of time periods as well as the different ways of summing up the sentiments. In one approach of summing up sentiments, the sentiments of multiple users are equally treated and added up to generate the collective sentiment; in the other approach, different users' unified sentiments are assigned with different weights, which are the users' numbers of followers (in the calculation, it is added by 1 to count the user himself/herself into the equation).

Granger Causality test is carried out to study whether prediction power exists between users' collective sentiments of tweets and the stock market price. Each pair of the series of the 16 stocks is studied individually. It is not to test the actual causation, but whether one time series has predictive information about the other. Granger Causality has been widely used in economics since 1960s. Its mathematical formulation is based on linear regression modeling of stochastic processes (Granger, 1969). The basic GC definition is quite simple. Suppose that there are two terms X_t and Y_t, and the first attempt is made to forecast X_{t+1} using past terms X_t, then the second attempt is made to forecast X_{t+1} using past terms of X_t and Y_t. If the second forecast is found to be more successful,

according to standard cost functions, then the past of Y appears to contain information helping in forecasting X_{t+1} that is not in past X_t. In particular, there could be other possible explanatory variables. Thus, Y_t would "Granger cause" X_t if (a) Y_t occurs before X_{t+1}; and (b) it contains information useful in forecasting X_{t+1} that is not found in a group of other appropriate variables. In the case of the thesis, if Y is the collective sentiment and X is the stock price change, a GC test can demonstrate whether the collective sentiment appears to contain information helping in forecasting the stock price change of tomorrow that is not in the stock price change of today.

If it turns out that majority of the time series pairs pass the Granger Causality test, a further step is taken to test the collective sentiments' prediction power on the direction of the stock price movement. Binary Logistic Regression is used to do the prediction. The results of the analyses are as follows.

6.2.1 Volume Correlation

Bivariate correlation analyses are conducted to study the correlations between the unified volume of tweets and the trading volume. The Pearson product-moment correlation coefficient (sometimes referred to as Pearson's r, and is typically denoted by r) (Stigler, 1989) is the measure of the correlation (linear dependence) between two variables X and Y, giving a value between +1 and −1 inclusive. It is widely used in the sciences as a measure of the strength of linear dependence between two variables. That formula for *r* is:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

From Table 8, there are significant positive correlations (with critical value of approximately 0.3) on the same day for all of the 16 stocks for the period of 12:00am – 11:59pm. For the time period of 4:00pm – 9:30am in Table 9, the significant positive correlations (with critical value of approximately 0.3) are found in 15 out of the16 stocks. Although the time stamps of the two time series are lined up for the period of 4:00pm-9:30am, the period during which the tweets are published is actually before the daily trading starts. Thus there is certain prediction power that can be seen from this correlation. It is valuable information that users' activity on StockTwits overnight significantly positively correlates to the stock trading volume the next business day. This can be explained by the relationship between the market capitalization and the daily trading volume, and the relationship between the daily trading volume and the public attention. Although it is not likely to be causality, it is still clear evidence that the user base on StockTwits is a decent representation of the public who engages in stock market.

Table 8: Result of Volume Correlation for 12:00am – 11:59pm

Stock	**$AAPL**	**$AMZN**	**$BAC**	**$BIDU**	**$C**	**$CMG**	**$FSLR**	**GOOG**
Correlation	0.871	0.855	0.936	0.858	0.792	0.830	$0.868	0.829
Stock	**$GS**	**$IBM**	**$JPM**	**$MSFT**	**$NFLX**	**$PCLN**	**$RIMM**	**$YHOO**
Correlation	0.73	0.591	0.834	0.543	0.851	0.729	0.957	0.390

Table 9: Result of Volume Correlation for 4:00pm – 9:30am

Stock	**$AAPL**	**$AMZN**	**$BAC**	**$BIDU**	**$C**	**$CMG**	**$FSLR**	**GOOG**
Correlation	0.530	0.566	0.602	0.757	0.654	0.745	0.760	0.454
Stock	**$GS**	**$IBM**	**$JPM**	**$MSFT**	**$NFLX**	**$PCLN**	**$RIMM**	**$YHOO**
Correlation	0.540	0.679	0.772	0.681	0.677	0.226	0.359	0.503

6.2.2 Sentiment and the Stock Market

Granger Causality is a statistical concept of causality that is based on prediction. According to Granger Causality, if a signal X1"Granger-causes" (or "G-causes") a signal X2, then past values of X1 should contain information that helps predict X2 above and beyond the information contained in past values of X2 alone. GC is normally tested in the context of linear regression models. For illustration, consider a bivariate linear autoregressive model of two variables $X1$ and $X2$:

$$X_1(t) = \sum_{j-1}^{p} A_{11,j} X_1(t-j) + \sum_{j-1}^{p} A_{12,j} X_2(t-j) + E_1(t)$$

$$X_2(t) = \sum_{j-1}^{p} A_{21,j} X_1(t-j) + \sum_{j-1}^{p} A_{22,j} X_2(t-j) + E_2(t)$$

where p is the maximum number of lagged observations included in the model, the matrix A contains the coefficients of the model (i.e., the contributions of each lagged observation to the predicted values of $X1(t)$ and $X2(t)$, and $E1$ and $E2$ are residuals (prediction errors) for each time series. If the variance of $E1$ (or $E2$) is reduced by the inclusion of the $X2$ (or $X1$) terms in the first (or second) equation, then it is said that $X2$ (or $X1$) Granger-(G)-causes $X1$ (or $X2$). In other words, $X2$ G-causes $X1$ if the coefficients in $A12$ are jointly significantly different from zero. This can be tested by performing an F-test of the null hypothesis that $A12 = 0$, given assumptions of covariance stationarity on $X1$ and $X2$ (Seth, 2007).

To apply Granger Causality test, each time series needs to be stationary. Although Bollen et al.'s method of stationarizing the series is a good approach, which is to normalize the series to z-scores on the basis of a local mean and standard deviation within a sliding

window of a certain number of days before and after the particular date(Bollen, Mao, & Zeng, 2011), it can only be used to study the causality but does not have practical application because one can never get data from the future to study the present. For that reason, Augmented Dickey Fuller (ADF) Unit Root Test of each series is examined individually to check its stationarity before the GC test. Normally, if the series cannot pass the ADF test, the first difference should be taken to stationarize the series. However, there is not much rational reasoning of taking the first difference in this dataset both due to its coarse nature (it is estimated from the social media) and the assumption that the day-to-day collective sentiments on the stock market should be random. If the series fails to pass the ADF test, it will be dropped from the analyses.

Six different types of collective sentiments are calculated and are tested on its stationarity by using ADF as described on all of the 16 stocks. Table 10 presents the list of the series which passed the stationarity test. Out of a total of 96 series, majority of the series passed the ADF test with 10 exemptions.

The series are then tested against the daily stock price change, which is defined as:

$$P_{i\_change} = \frac{P_i - P_{i-1}}{P_{i-1}}.$$

The test is conducted at the lag of 1 (p is 1) because the volume correlation reveals that users' activity on StockTwits overnight significantly positively correlates to the stock trading volume the next business day, so the relationship may also exist for the collective sentiments and the stock price change. A total number of 86 one-to-one Granger Causality tests are systematically carried out to study the relationship between the collective sentiments and the change of stock. The following Table 11 presents a

summarization of the Granger Causality test results. Graphs of some series are also analyzed for further information.

Table 10: Results of ADF Test

Stock	Full Day Simple Sum	Afterhours Simple Sum	Full Day Simple Sum Normalized	Afterhours Simple Sum Normalized	Full Day Weighed Sum	Afterhours Weighed Sum
$AAPL			x	x	x	x
$AMZN		x			x	x
$BAC		x	x	x	x	x
$BIDU		x	x	x	x	x
$C	x	x	x	x	x	x
$CMG	x	x	x	x	x	x
$FSLR	x	x	x	x	x	x
$GOOG		x		x	x	x
$GS	x	x	x	x	x	x
$IBM	x	x	x	x	x	x
$JPM	x	x	x		x	x
$MSFT	x	x	x	x	x	x
$NFLX	x	x	x	x	x	x
$PCLN	x	x	x	x	x	x
$RIMM	x	x	x	x	x	x
$YHOO	x	x	x	x	x	x

- **Full Day**: 12:00am-11:59pm
- **Afterhours**: 4:00pm- 9:30am
- **Simple Sum**: Simply summed up collective sentiments
- **Weighed Sum**: Weighed summed up collective sentiments
- **Simple Sum Normalized:** Simply summed up collective sentiments divided by the unified volume of tweets during the same period

Table 11: Results of Granger Causality Test

Stock	Full Day Simple Sum	Afterhours Simple Sum	Full Day Simple Sum Normalized	Afterhours Simple Sum Normalized	Full Day Weighed Sum	Afterhours Weighed Sum
$AAPL			x	x	x	x
$AMZN		$S \xrightarrow{at\ 0.05} C$			x	x
$BAC		$S \xrightarrow{at\ 0.05} C$ $C \xrightarrow{at\ 0.05} S$	x	x	$S \xrightarrow{at\ 0.10} C$	$C \xrightarrow{at\ 0.05} S$
$BIDU		x	x	x	$S \xrightarrow{at\ 0.05} C$	x
$C	x	x	x	x	$S \xrightarrow{at\ 0.05} C$	x
$CMG	x $C \xrightarrow{at\ 0.05} S$	$S \xrightarrow{at\ 0.05} C$	x $C \xrightarrow{at\ 0.05} S$	x	x	x
$FSLR	x	$S \xrightarrow{at\ 0.05} C$	x	x	x	$S \xrightarrow{at\ 0.05} C$
$GOOG		x		x	x	x
$GS	x $C \xrightarrow{at\ 0.1} S$	x $C \xrightarrow{at\ 0.1} S$	x	x	x $C \xrightarrow{at\ 0.05} S$	x $C \xrightarrow{at\ 0.05} S$
$IBM	$S \xrightarrow{at\ 0.05} C$	$S \xrightarrow{at\ 0.05} C$	x	x	x	x
$JPM	x $C \xrightarrow{at\ 0.05} S$	x $C \xrightarrow{at\ 0.05} S$	x		x	x
$MSFT	$S \xrightarrow{at\ 0.05} C$	$S \xrightarrow{at\ 0.05} C$	x	x	x	x
$NFLX	x	$S \xrightarrow{at\ 0.05} C$	x	$C \xrightarrow{at\ 0.1} S$	x	$S \xrightarrow{at\ 0.05} C$
$PCLN	x	x $C \xrightarrow{at\ 0.1} S$	x	x	x	x
$RIMM	x	$S \xrightarrow{at\ 0.05} C$	x	x	x	x
$YHOO	x	$S \xrightarrow{at\ 0.05} C$	x	x	x	$S \xrightarrow{at\ 0.05} C$ $C \xrightarrow{at\ 0.05} S$

- **x**: No significant relationship
- **C**: Change of stock
- **S**: Sentiment
- **at 0.05, at 0.1**: Probability of accepting the Null Hypothesis

For the collective sentiments of "Full Day Simple Sum", collective sentiments G-cause change of stock in 2 out of 11 tested stocks; and change of stock G-causes collective sentiments in another 3 out of 11 tested stocks. For "Full Day Weighed Sum", collective sentiments G-cause change of stock in 3 out of 16 stocks; and change of stock G-causes collective sentiments in 3 out of 16 stocks as well. For "Afterhours Simple Sum", collective sentiments G-cause change of stock in 9 out of 15 tested stocks; and change of stock G-causes collective sentiments in 4 out of 15 stocks. For "Afterhours Weighed Sum", collective sentiments G-cause change of stock in 3 out of 16 stocks, and change of stock G-causes collective sentiments in 1 out of 16 stocks. The numbers are 0 out of 14 and 1 out of 14 for both "Full Day Simple Sum Normalized" and "Afterhours Simple Sum Normalized".

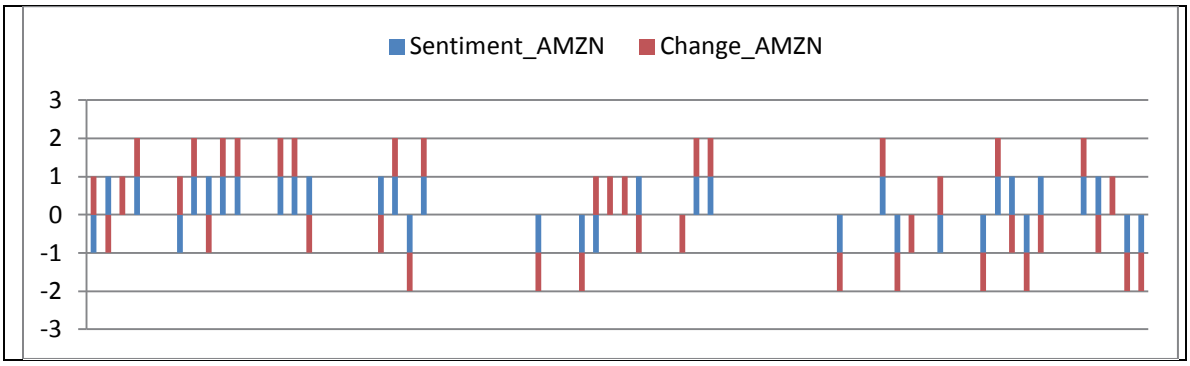
On one hand, the findings on the fact that collective sentiments G-cause change of stock are exciting. The "Afterhours Simple Sum", which is the simply summed up collective sentiments for afterhours, has powerful prediction on the change of stock price for the next day in most of the stocks studied. When the same collective sentiments are normalized by the numbers of users, however, the prediction power becomes dissolved by the noises generated by having too many users count into the equation. Similarly, when the unified sentiments are assigned with weights, which are the numbers of the publishers' followers, the prediction power is not as strong as simply summed up sentiments. On the other hand, the reverse process that changes of stock G-causes collective sentiments is also found in several stocks by taking different approaches. Interestingly, the one with the most relationships is again the "Afterhours Simple Sum" approach of generating collective sentiments; and 3 out the 4 relationships are in

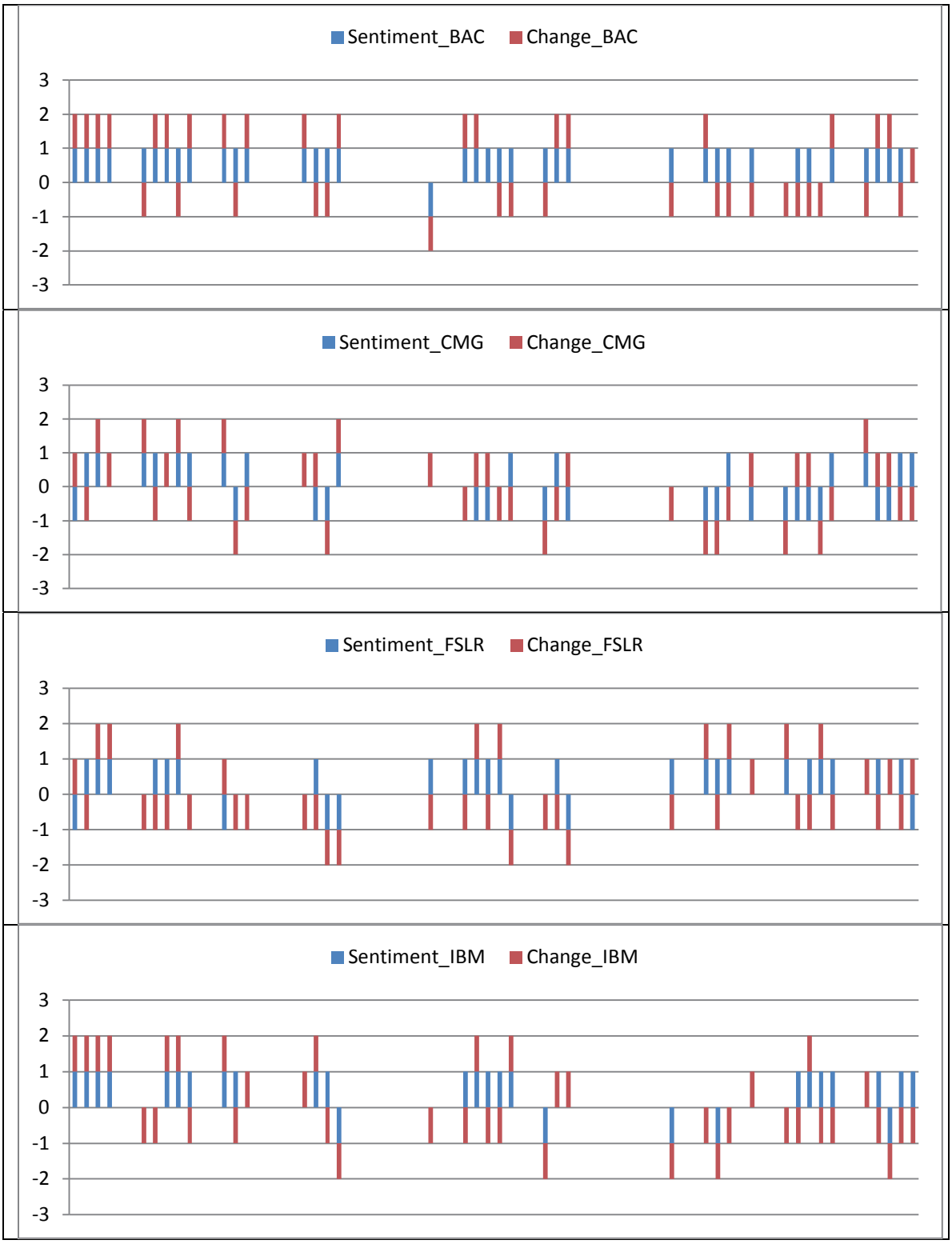
completely different stocks. It is profound finding that the Granger Causality exists in different directions in different stocks. For some stocks, the collective sentiments have prediction power on the stock price change for the next day (The time series of 9 stocks whose sentiments G-cause change of stocks are presented in Appendix A: Time Series of 9 Stocks Whose Sentiments G-cause Changes of Stocks); while for some other stocks, the stock price change actually influences users' collective sentiments for the next day (Appendix B: Time Series of 4 Stocks Whose Changes of Stocks G-cause Sentiments presents the time series plots in which collective sentiments are shifted for one day in the reverse direction).

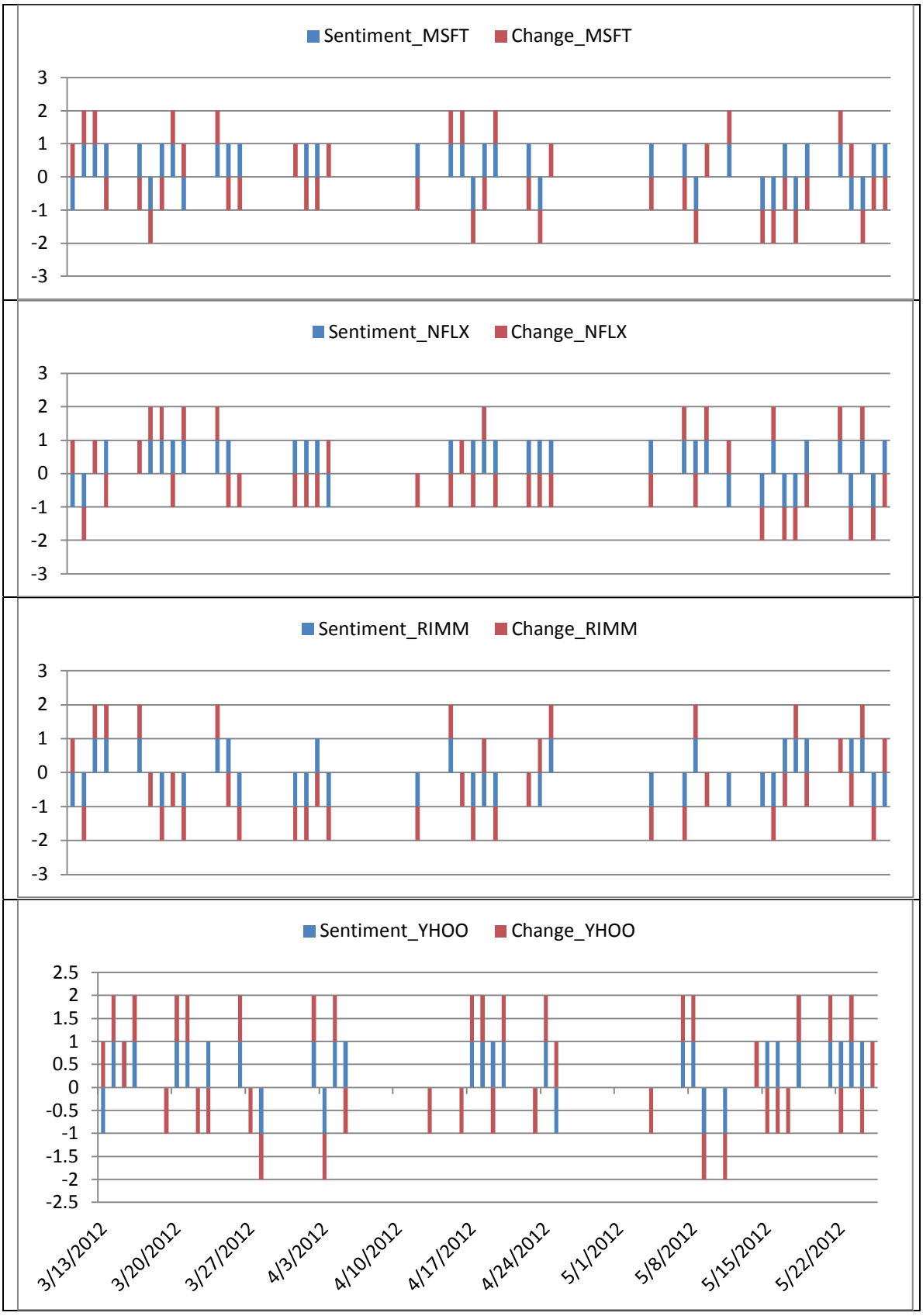
6.2.3 Sentiment and the Direction of the Stock Movement

Figure 7 presents the 9 pairs of the bi-directional movement plots of the 9 stocks without considering the magnitude of the movement. "Afterhours Simple Sum" collective sentiments are used to predict the stock movement for the next day. The time series for the sentiment have been shifted for one day on the chart. The co-movements of the two series can be noticed in most of the series. For example, the bi-directional co-movement can be seen 72.5% times on the $RIMM time series, and 70% of times on the $YHOO time series.

Figure 7: Bi-directional Plots of Sentiments and Change of Stock




Sentiment_BAC
Change_BAC
3
2
1
0
-1
-2
-3
Sentiment_CMG
Change_CMG
3
2
1
0
-1
-2
-3
Sentiment_FSLR
Change_FSLR
3
2
1
0
-1
-2
-3
Sentiment_IBM
Change_IBM
3
2
1
0
-1
-2
-3


Sentiment_MSFT
Change_MSFT
3
2
1
0
-1
-2
-3
Sentiment_NFLX
Change_NFLX
3
2
1
0
-1
-2
-3
Sentiment_RIMM
Change_RIMM
3
2
1
0
-1
-2
-3
Sentiment_YHOO
Change_YHOO
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
3/13/2012
3/20/2012
3/27/2012
4/3/2012
4/10/2012
4/17/2012
4/24/2012
5/1/2012
5/8/2012
5/15/2012
5/22/2012

If a simple model is developed by classifying all the negative collective sentiments to cause negative movement (and vice versa), the overall prediction accuracy of the 9 stock is 54.4%. The prediction accuracy of negative stock movement is 60.8% while the one for positive movement is 49.8%. The details can be found in Table 12.

Table 12: Classification Table of Simple Stock Direction Prediction

Classification Table

Observed		Predicted		
		Change_Stock (Binned)		Percentage Correct
		Fall	Rise	
Change_Stock (Binned)	Fall	90	104	46.4
	Rise	58	103	64
Overall Percentage		60.8	49.8	54.4

To further test the prediction accuracy of the up and down movement of the stock prices, Binary Logistic Regression is carried out by using the variables "Afterhours Simple Sum" sentiments and the afterhours unified volumes of tweets before the market opens. The results are in Table 13. The overall accuracy is increased to 58.9%. The accuracy rate of predicting that the stock price will fall and the accuracy rate of predicting the opposite are both 58.9%.

Table 13: Classification Table of Stock Direction Prediction

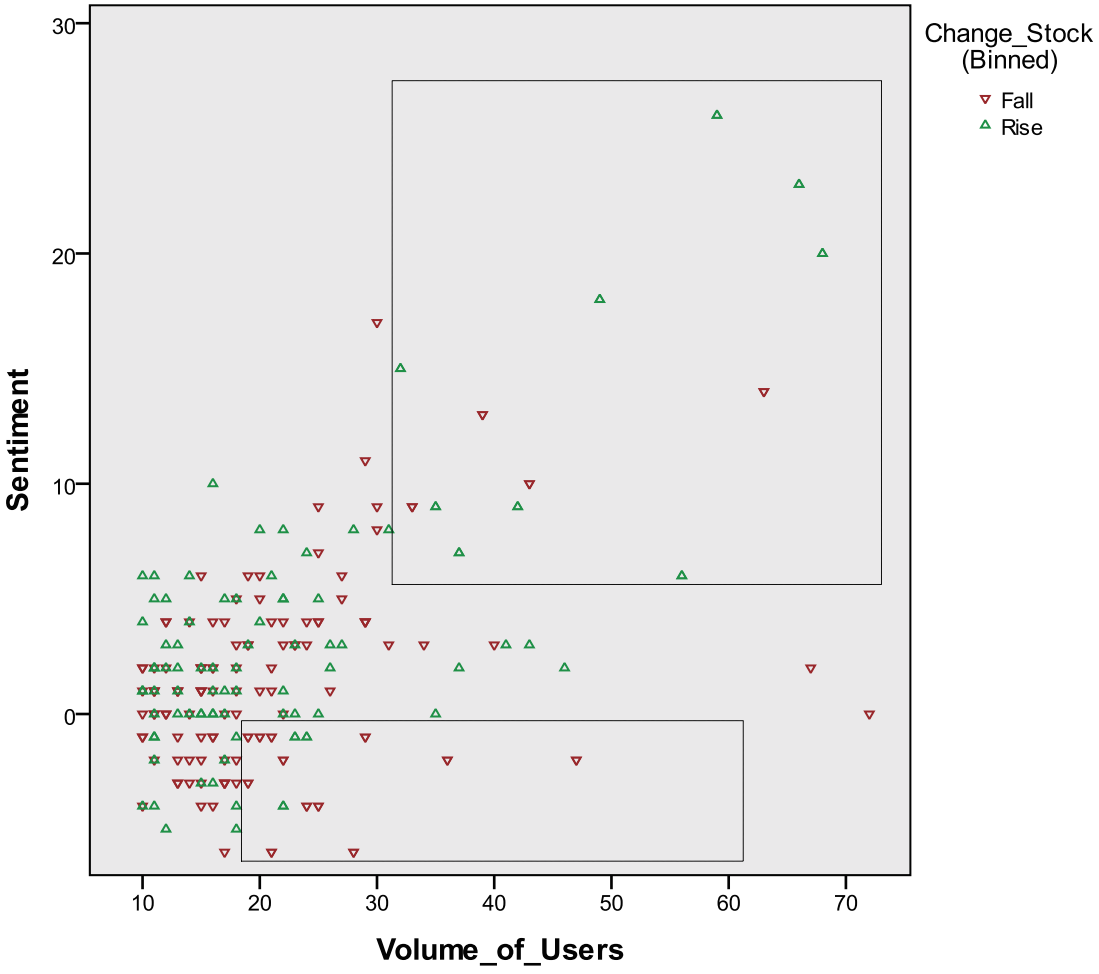
Classification Table

Observed		Predicted		
		Change_Stock (Binned)		Percentage Correct
		Fall	Rise	
Change_Stock (Binned)	Fall	169	28	85.8
	Rise	118	40	25.3
Overall Percentage		58.9	58.9	58.9

The cut value is .500

Figure 8 presents the scatter plot of bi-directional stock movements by using "Afterhours Simple Sum" sentiments and the afterhours unified volumes of tweets before the market opens. It seems that the two variables are better in predicting negative stock movements. When the sentiment is negative, most of the stock movements are negative. When the sentiment is negative and the volume is large, the prediction power of negative stock movements becomes even stronger. Similar observations can be found for positive stock movements when the sentiment is positive and the volume is large. However, the data is too scarce to draw a conclusion.

Figure 8: Scatter Plot of Bi-directional Stock Prediction



To test whether the accuracy rate can be improved by taking into account the magnitude of the up and down movement of the stock prices, the target variable is then refined to exclude the medium range of the stock price change, which is between -1% and 1%. The data size is thus reduced and the attributes between the up and down movement of the stock prices are believed to be more distinguished in this case. However, Table 14 shows that the overall prediction accuracy cannot be improved. Although the accuracy rate of predicting the stock will fall is improved to be 59.8%, the accuracy of the opposite is reduced to be 54.5%. It is worth noting that in both of the logistic regression models, the change of stock are classified to fall for more than 80% of the times, while in reality it should normally be 50%.

Table 14: Classification Table of Stock Direction Prediction Refined

Classification Table

Observed		Predicted		
		Change_Stock (Binned)		Percentage Correct
		Big Fall	Big Rise	
Change_Stock (Binned)	Big Fall	98	15	86.7
	Big Rise	66	18	21.4
Overall Percentage		59.8	54.5	58.9

The cut value is .500

Chapter 7 CONCLUSION AND FUTURE WORK

The stock market prediction has been intensively studied by scholars and professionals from finance domain, information management domain and even psychology domain. Social interaction is an important aspect of the decision-making process. People obtain information and opinions about a decision by communicating with one another. A positive or negative sentiment is quickly reflected in the stock market. Online social media, a novel form of communication emerged over the past five years, has just started gaining enough data for carrying out analysis.

This thesis has empirical contributions to this research area by taking a unique approach of public sentiment analysis and also has practical implications by proposing an effective technique of stock market analysis.

7.1 EMPIRICAL CONTRIBUTIONS

Previous work in this area of sentiment analysis traditionally focuses on product reviews by using lexicon-based approach, usually counting the positive and negative polarity words. In this thesis, machine learning algorithms are used in natural language processing to get the public sentiment on individual stocks in order to study its relationship with the stock price change.

The research architecture consists of a NLP approach and a statistical analysis approach. The NLP approach of sentiment detection is again a two-stage process by implementing Neutral v.s. Polarized detection before Positive v.s. Negative detection. The two-stage approach is in line with Wilson et al.'s research (Wilson, Wiebe, & Hoffmann, 2009) on how the presence of neutral instances may affect the performance of features for

distinguishing between positive and negative polarity. The statistical approach takes a unique path in dealing with the time in light of the thought that the sentiments during the open time of the market may be influenced by the real-time market fluctuations thus create noises in the sentiments' prediction power for the future. In one case, the sentiments of the tweets posted during 12:00am and 11:59pm on the same day are aggregated, while in the other case, the sentiments of tweets posted during the period of 4:00pm (the marketing closing time) and the next day 9:30am (the market opening time) are aggregated. It is proved that the collective sentiments of afterhours are much stronger predictors.

In the thesis, the initial assumption that the users on StockTwits have the genuine incentive to produce high-quality content is validated through the hand-labeling process as well as the proven prediction power. Besides, attempt has also been made in the thesis on weighing the sentiments of different users by putting into the equation the measurements of experts' followers to reflect their public influence. However, the prediction power is not as strong as simply summed up sentiments. This could be due to the over-simplified method of weighing the unified sentiments with the number of followers.

7.2 PRACTICAL IMPLICATIONS

The objectives of the thesis have been achieved by having:

- Demonstrated that Support Vector Machine is the best classifier with the overall accuracy rates of 71.84% and 74.3%, respectively, at the two-stage sentiment detection process.

- Discovered that users' activity on StockTwits overnight significantly positively correlates to the stock trading volume the next business day.
- Determined that simply summed up collective sentiments for afterhours has powerful prediction on the change of stock price for the next day in 9/15 of the stocks studied by using Granger Causality test; discovered that the overall accuracy rate of predicting the up and down movement of stocks by using the collective sentiments is 58.9%.

The overall accuracy rates of 71.84% and 74.3% are satisfying results, as the process of determining the sentiment of a tweet is vague even for human and people only agree on sentiment 80% of the time. The fact that users' activity on StockTwits overnight significantly positively correlates to the stock trading volume the next business day is clear evidence that the user base on StockTwits is a decent representation of the public who engages in stock market.It can also be concluded from the analyses that the collective sentiments of afterhours have certain prediction power on the direction of the stock movement of the next day.

7.3 FUTURE WORK

There are certain limitations in the thesis. First of all, the data has been collected only over a period of two and a half months. If data can be collected over a longer period, the result is believed to be more significant. Secondly, the knowledge corpus required for understanding the financial sentiments evoked by words and phrases is huge. An expanded lexicon from the training data might help increase the classification accuracy rate in sentiment detection. Thirdly, the detailed user profile information such as level of

experience, approach of trade or holding period of trade is not used due to the data scarcity. If the holding period of trade is taken into account, for example, day traders can be grouped to predict intraday or day-to-day stock price change, short term traders can be grouped to predict week-to-week stock price change, etc., it is believed that better result could be achieved.

Moreover, further research can be done by separating companies by types. It is already discovered from the thesis that the directions in which the Granger Causality is found are different for different types of stocks. In Ruiz et al.'s research (Ruiz & Hristidis, 2012) on correlating financial time series with micro-blogging activity, where they find that the correlation is stronger for companies with low debt, regardless of whether their financial indicators are healthy or not, they also find that the users' tweets correlate better with the stocks for companies having high beta (A brief description of beta can be found in Appendix C: A Brief Description of Beta), which implies a stock price grows dramatically when the market is up, and falls dramatically when the market goes down. Small values of beta mean the stock's return is relatively unaffected by the swings in the overall market's return, again suggesting that Twitter activity seems to be better correlated with traded volume for companies whose finances fluctuate a lot. It is believed that if similar research can be done on the thesis' set of data, there will be similar findings.

Apart from what have been mentioned above, it is also suggested that for future research in similar topics approaches can be taken by weighing the sentiments based on users' historical prediction accuracy, i.e. giving more weight to users with high accuracy rates in the past. Alternatively, network features such as the number of connected points

regarding a particular stock discussion can be extracted from the data to be included in the analysis.

Last but not least, it has been observed from much literature that simulations are carried out by including the new features generated. A simple simulation does not have to consider external effects like deficit of stocks. It can be used to determine if the proposed tweet features have the potential of improving over the other baseline strategies.

BIBLIOGRAPHY

Aggarwal, S., & Aggarwal, N. (2011). Classification of Audio Data using Support Vector Machine. *International Journal of Computer Science and Technology*, IJCST Vol. 2, Issue 3, September.

Avery, C., Chevalier, J. A., & Zeckhauser, R. J. (2011, August). The "CAPS" Prediction System And Stock Market Returns. *NBER Working Paper Series*. Cambridge, MA, USA: National Bureau Of Economic Research.

Babu, M., N.Geethanjali., & Kumari, V. (2010). Texual Analysis of Stock Market Prediction using Financial News Articles. *The Technology World Quarterly Journal*, December Volume II Issue 4.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. *7th Conference on Language Resources and Evaluation* (pp. 2200-2204). Valletta: In Proceedings of LREC-10.

Bakshy, E., & Hofman, J. M. (2011). Everyone's an Influencer:Quantifying Influence on Twitter. *WSDM* (pp. 65-74). Hong Kong: ACM 978-1-4503-0493-1/11/02.

Balthasar, A. (2009). *Using Google Search Volume as Trading Signal.* London: Barclays Bank PLC.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter Mood Predicts the Stock Market. *Journal of Computational Science*, 1-8.

Butler, M. (2009). *An Artificial Intelligence Approach to Financial Forecasting using Improved Data Representation, Multi-objective Optimization, and Text Mining.* Halifax: Dalhousie University Thesis.

Butler, M., & Keselj, V. (2009). Financial Forecasting using Character N-Gram Analysis and Readability Scores of Annual Reports.

Butler, M., & Keselj, V. (2009). Optimizing a Pseudo Financial Factor Model with Support Vector Machine and Genetic Programming.

Chen, H., Prabuddha, D., Hu, Y., & Hwang, B.-H. (2011). Sentiment Revealed in Social Media and its Effect on the Stock Market. *IEEE Statistical Signal Processing Workshop*, 25-28.

Cheung, M. Y., Luo, C., Sia, C. L., & Chen, H. (Summer 2009). Credibility of Electronic Word-of-Mouth:Informational and Normative Determinants of On-line Consumer Recommendations. *International Journal of Electronic Commerce*, Vol. 13, No. 4, p9-38.

Choudhury, M. D., Sundaram, H., John, A., & Seligmann, D. D. (2008, June 19–21). Can Blog Communication Dynamics be Correlated with Stock Market Activity? *HT'08*. Pittsburgh, Pennsylvania, USA.

Dinu, L. P., & Iuga, I. (2012). The Naive Bayes Classifier in Opinion Mining: In Search of the Best Feature Set. *Lecture Notes in Computer Science*, Volime 7142/2012, 210-222.

Fama, E. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *Journal of Finance*, XXV,No.2,pages 226-241.

Golub, B., & Jackson, M. O. (2010). Naive Learning in Social Networks and the Wisdom of Crowds. *American Economic Journal: Microeconomics*, 2:1, 112-149.

Granger, C. W. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *The Econometric Society*, 424-438 .

Grimes, S. (2010, March 29). *Expert Analysis: Is Sentiment Analysis an 80% Solution?* Retrieved July 22, 2012, from InformationWeek: http://www.informationweek.com/news/224200667

Gu, B., Konana, P., Liu, A., Rajagopalan, B., & Ghosh, J. (2006, November). Predictive Value of Stock Message Board Sentiments. *McCombs Research Paper*.

Guo, L., Tan, E., Chen, S., Zhang, X., & Zhao, Y. (2009, June 28–July 1). Analyzing Patterns of User Content Generation in Online Social Networks. Paris, France.

Hill, S., & Ready-Campbell, N. (2011). Expert Stock Picker: The wisdom of (Experts in) Crowds. *International Journal of Electronic Commerce*, Vol. 15, No. 3, p73-101.

Investopedia. (n.d.). Retrieved July 16, 2012, from http://www.investopedia.com/exam-guide/cfa-level-1/derivatives/options-calls-puts.asp#axzz20osP1hRC

Investopedia. (n.d.). *Beta*. Retrieved July 20, 2012, from Investopedia: http://www.investopedia.com/terms/b/beta.asp#axzz22S0tj0Ec

Lazarsfeld, P., Berelson, B., & Gaudet, H. (1944). The People's Choice: How the Voter Makes up His Mind in a Presidential Campaign. *New York: Columbia University Press*.

Lindzon, H., Pearlman, P., & Ivanhoff, I. (2011). *The StockTwits Edge.* New Jersey: John Wiley & Sons.

Loughran, T., & McDonald, B. (2011). When is a Liability not a Liability? *The Journal of Finance*.

Nofsinger, J. R. (2005). Social Mood and Financial Economics. *The Journal of Behavioral Finance*, 144-160.

Pajupuu, H., Kerge, K., & Rene. (2011). Lexicon-based Detection of Emotion in Different Types of Texts: Preliminary Remarks. *EESTI Rakenduslingvistika Ühingu Aastaraamat*, 171-184.

Pang, B., & Lee, L. (2004). A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. *ACL-04, 42nd Meeting of the Association for Computational Linguistics*, (pp. 271-278). Barcelona.

Pang, B., & Lee, L. (2005). Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. *ACL-05, 43rd Meeting of the Association for Computational Linguistics*, (pp. 115–124). Ann Arbor.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. *EMNLP* (pp. 79--86). Philadelphia: Association for Computational Linguistics.

Rosé, C., Wang, Y.-C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., & Fischer, F. (2007). Analyzing Collaborative Learning Processes Automatically:Exploiting the Advances of Computational Linguistics in Computer-supported Collaborative Learning. *International Journal of Computer Supported Collaborative Learning*.

Ruiz, E. J., & Hristidis, V. (2012). Correlating Financial Time Series with Micro-Blogging Activity. *Web Search and Data Mining (WSDM)*. Seattle.

Satapathy, S., & Bhagwani, S. (2012). *Capturing Emotions in Sentences.* Retrieved March 15, 2012, from http://202.3.77.10/users/sranjans/emotionDetection.pdf

Schumaker, R., Zhang, Y., & Huang, C. (2011). Sentiment Analysis of Financial News Articles. *Communications of the International Information Management Association*, forthcoming.

Seth, A. (2007). *Granger causality - Scholarpedia 2(7):1667., revision #91329.* Retrieved July 16, 2012, from Scholarpedia: http://www.scholarpedia.org/article/Granger_causality

Stigler, S. M. (1989). Francis Galton's Account of the Invention of Correlation. *Statistical Science*, 4 (2): 73–79.

Wikipedia. (2012, May 25). *Social trading*. Retrieved July 20, 2012, from Wikipedia: http://en.wikipedia.org/wiki/Social_trading

Wikipedia. (n.d.). *Beta (finance)*. Retrieved July 20, 2012, from Wikipedia: http://en.wikipedia.org/wiki/Beta_(finance)

Williams, R. T. (2011). *An Introduction to Trading in the Financial Markets.* Elsevier.

Wilson, T., Wiebe, J., & Hoffmann, P. (2009). Recognizing Contextual Polarity:An Exploration of Features for Phrase-Level Sentiment Analysis. *Computational Linguistics*, 399-433, Volume 35, Number 3.

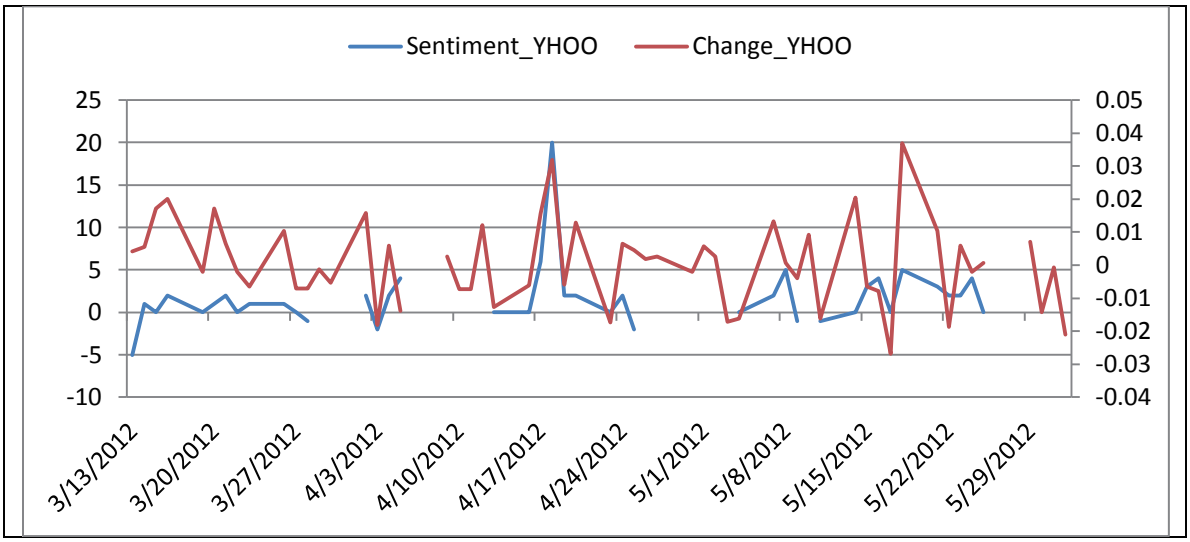
World Finance on Social Trading. (n.d.). Retrieved July 20, 2012, from World Finance: http://www.social-trading.worldfinance.com/

Wuthrich, B., Cho, V., Leung, S., Permunetilleke, D., Sankaran, K., Zhang, J., & Lam, W. (1998). Daily Stock Market Forecast from Textual Web Data. *IEEE International Conference On Systems, Man And Cybernetics*, (pp. 2720--2725). Hong Kong.

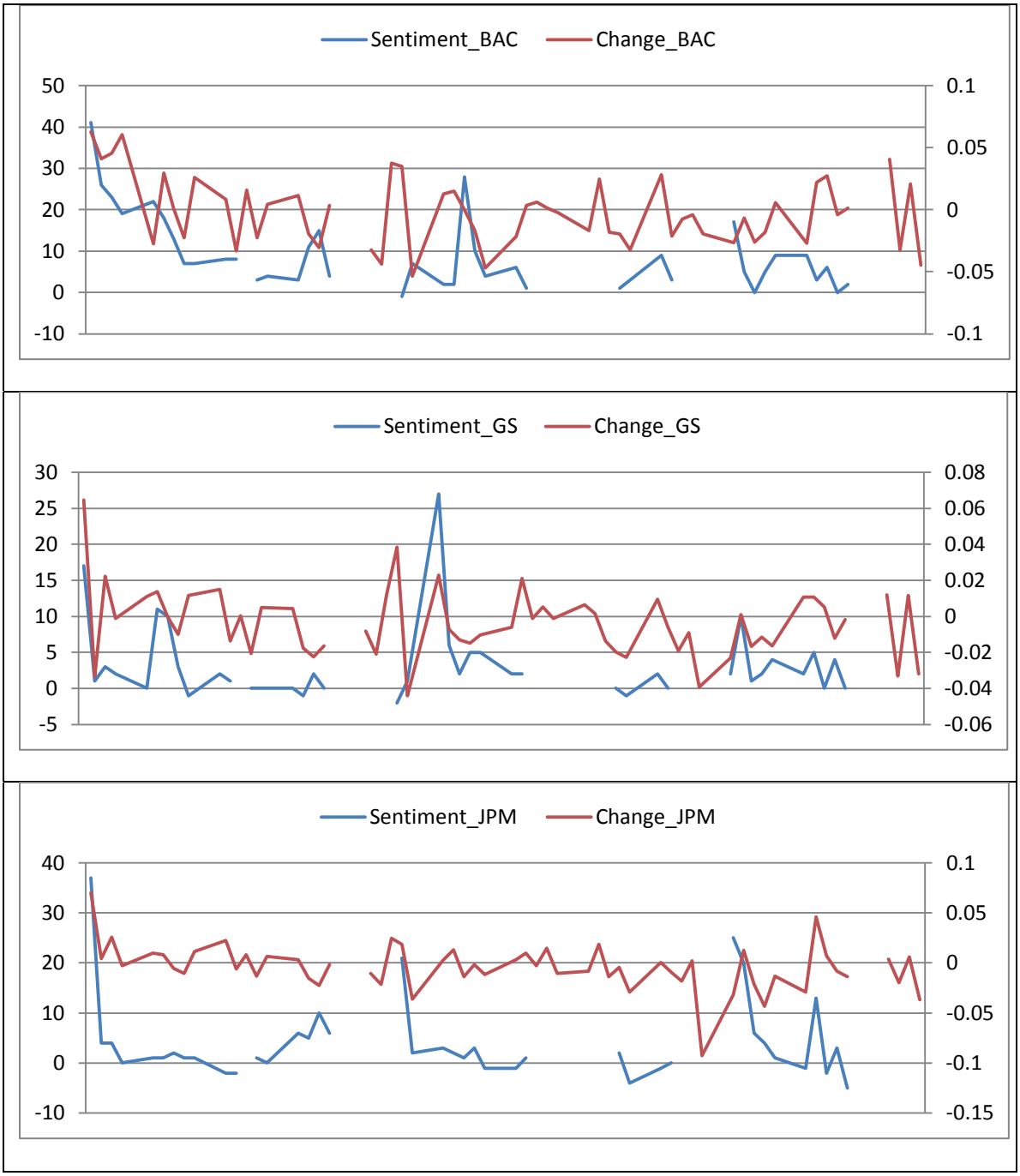
Zhai, J., Cohen, N., & Atreya, A. (2011). *Sentiment Analysis of News Articles for Financial Signal Prediction.* Standford Student Course Work.

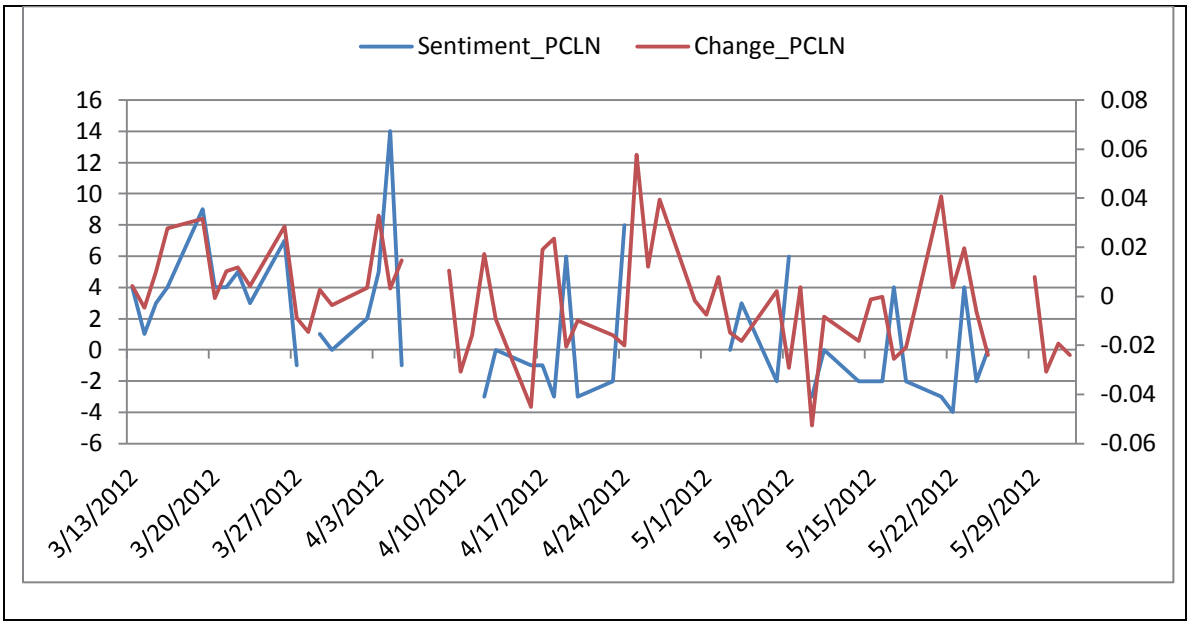
APPENDIX A: TIME SERIES OF 9 STOCKS WHOSE SENTIMENTS G-CAUSE CHANGES OF STOCKS


Sentiment_IBM
Change_IBM
40
30
20
10
0
-10
0.04
0.02
0
-0.02
-0.04
Sentiment_MSFT
Change_MSFT
40
30
20
10
0
-10
0.06
0.04
0.02
0
-0.02
-0.04
Sentiment_NFLX
Change_NFLX
30
20
10
0
-10
0.1
0.05
0
-0.05
-0.1
-0.15
Sentiment_RIMM
Change_RIMM
10
5
0
-5
-10
0.1
0.05
0
-0.05
-0.1
-0.15


Sentiment_YHOO
Change_YHOO
25
20
15
10
5
0
-5
-10
0.05
0.04
0.03
0.02
0.01
0
-0.01
-0.02
-0.03
-0.04
3/13/2012
3/20/2012
3/27/2012
4/3/2012
4/10/2012
4/17/2012
4/24/2012
5/1/2012
5/8/2012
5/15/2012
5/22/2012
5/29/2012

APPENDIX B: TIME SERIES OF 4 STOCKS WHOSE CHANGES OF STOCKS G-CAUSE SENTIMENTS




Sentiment_PCLN
Change_PCLN
16
14
12
10
8
6
4
2
0
-2
-4
-6
0.08
0.06
0.04
0.02
0
-0.02
-0.04
-0.06
3/13/2012
3/20/2012
3/27/2012
4/3/2012
4/10/2012
4/17/2012
4/24/2012
5/1/2012
5/8/2012
5/15/2012
5/22/2012
5/29/2012

APPENDIX C: A BRIEF DESCRIPTION OF BETA

In finance, the Beta (β) of a stock or portfolio is a number describing the volatility of an asset in relation to the volatility of the benchmark that said asset is being compared to. This benchmark is generally the overall financial market and is often estimated via the use of representative indices, such as the S&P 500 (Wikipedia).

Beta is calculated using regression analysis, and it can be described as the tendency of a security's returns to respond to swings in the market. A beta of 1 indicates that the security's price will move with the market. A beta of less than 1 means that the security will be less volatile than the market. A beta of greater than 1 indicates that the security's price will be more volatile than the market. A beta of 0 means that the security's returns change independently of changes in the market's returns. In general, a positive beta means that the asset's returns generally follow the market's returns, in the sense that they both tend to be above their respective averages together, or both tend to be below their respective averages together. A negative beta means that the asset's returns generally move opposite the market's returns: one will tend to be above its average when the other is below its average.

Many utilities stocks have a beta of less than 1. Conversely, most high-tech NASDAQ-based stocks have a beta of greater than 1, offering the possibility of a higher rate of return, but also posing more risk(Investopedia).