Ying Han. Correlating and Predicting Stock Prices with Twitter Sentiments. A Master's Paper for the M.S. in I.S degree. July, 2013. 44 pages. Advisor: Jaime Arguello

This paper presents an empirical study of correlating Twitter sentiments with individual stock price movements. We used an existing text-mining technique, OpinionFinder, to extract Twitter sentiment data from plaintext tweets. Different from prior researches, we explored a novel approach to aggregate Twitter sentiment features and Twitter metadata features associated with the tweets that mention a technology stock to construct a set of features, which was then correlated with the stock price movements of the respective stock prices. We thereby selected a subset of these features, which have positive correlation coefficients with the stock prices, to predict future stock price movements. The results of the prediction, however, are not as successful as expected. Although it is too early to conclude that Twitter sentiments cannot be used to predict an individual stock price, our results do provide one piece of negative evidence for such hypothesis.

Headings:

Machine learning

Text mining

Tweets (Microblogs)

CORRELATING AND PREDICTING STOCK PRICES WITH TWITTER SENTIMENTS

By
Ying Han

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

July 2013

Approved by

_____

Jaime Arguello

Table of Contents

1. Introduction

Stock market prediction has been an intriguing topic in both the real business world and the academic research. Early studies on stock market prediction based on random walk and Efficient Market Hypothesis (EMH) suggested that the stock market is unpredictable (Fama, 1965; Malkiel, 1973). However, recent researches on correlating events on social media with stock market movements have shown positive results. Especially in recent years, with the emergence and large-scale adoption of the real-time micro-blogging service, Twitter, people started to realize the information contained in Twitter tweets may have even better prediction power to the stock markets. In the world of academic research, it has been shown by scholars that Twitter data is positively correlated with stock trading prices or trading volumes (Bollen, Mao, & Zeng, 2010; Mao, Wang, Wei, & Liu, 2012; Ruiz, Hristidis, Castillo, Gionis, & Jaimes, 2012; Sprenger & Welpe, 2010; Yi, 2009; Zhang, Fuehres, & Gloor, 2011).

In this paper, we continue the research of the correlation between Twitter data and stock prices and trading volume. We share similar visions as previous studies that sentiments expressed in Twitter tweets can reflect to certain extend the public opinion towards the stock market, and hence the Twitter sentiments can be used for stock market movement correlation or even price prediction (Bollen et al., 2010; Zhang et al., 2011). Similar to Bollen et al. (2010), we used OpinionFinder (Wilson et al., 2005) to determine whether a tweet has positive sentiments or negative sentiments embedded in its text in an

automated way. However, unlike many previous studies, where the sentiment information is used to correlate and predict the stock market as a whole, our hypothesis is that the sentiments in the tweets mentioning a certain stocks will especially be able to correlate with the price movement of these individual stocks. As such, our study is different in the research goals from many previous researches.

Furthermore, in addition to sentiment information, we also utilized Twitter metadata in our correlation analysis. Particularly, we hypothesize that the metadata associated with the sentimental tweets, the tweets that contain explicit positive or negative user opinions, and the users who post the tweets may strengthen such correlation. Examples of these metadata include the total number of positive and negative tweets within a certain amount of time (such as an hour in our study), the total number of followers of the users who posted the positive (or negative) tweets, and the history of the Twitter users which implies the impacts of these users on other Twitter users. Our reasoning behind these is that, taking the number of followers of a Twitter user as an example, we conjecture that the greater the number is the more Twitter users are potentially influenced by the sentiments shared by the tweets. When combined with the Twitter metadata, the prediction power of each sentimental tweet is multiplexed by the potential influence it may have. Thus the total amount of positive or negative opinions reflected by the micro-blogs posted by the Twitter users will be able to provide a snapshot of the entire public opinion on the stock market and influence stock prices in the future. Consequently, by combining sentiment analysis with Twitter metadata, our approach is fundamentally distinct from previous methods which merely dealt with either

Twitter metadata features or sentiment features (Bollen et al., 2010; Mao et al., 2012; Ruiz et al., 2012; Yi, 2009; Zhang et al., 2011).

Therefore, the major contribution of this paper is that we propose a novel approach to combine sentiment features in Twitter tweets with features extracted from Twitter metadata for stock market movement correlation and prediction. The approach is comprised of extracting tweet sentiment features and aggregating Twitter metadata features. Our first step was to extract the sentiment features of tweets using OpinionFinder (Wilson et al., 2005). This text-based data-mining task is conducted by automatically identifying opinion sentiments and speculations in the text of the tweets. Then we aggregate the Twitter metadata collected together with the Twitter tweets to construct a set of metadata features and used SPSS to select the metadata features which have strong correlations with price directions of the stocks. The combined techniques provide a way to integrate the structured data (Twitter metadata) and unstructured data (Twitter tweet sentiments) for stock price correlation.

A second contribution of this paper is the results of our experiment evaluations. We show in our evaluation that some aggregated metadata features are more relevant to stock price changes while some are not. The existence of the correlations between these Twitter features and stock prices (and trading volumes) confirms that there exists some relation between Twitter sentiments and stock prices or volumes. However, our evaluation on using these positively correlated features to predict stock prices was not as successful as expected. Despite of the unsuccessful prediction, as an attempt to use combined Twitter sentiment data and metadata to predict stock prices, our study still shed light on the (in)effectiveness of such attempts, providing a piece of negative evidence to

the hypothesis that Twitter sentiments, when combined with Twitter metadata, can be used to predict individual stock prices.

The rest chapters of this thesis are structured as follows. Chapter 2 introduces the questions this thesis aims to research. A more detailed background on related work and background knowledge are given in Chapter 3. Then Chapter 4 presents Twitter and tweets data. The experimental design is outlined in Chapter 5. Chapter 6 presents how the Twitter features are selected, which is followed by Chapter 7, presenting the experiment results. We finally conclude our study in Chapter 8.

2. Research Question

While exploring the sentiment features embedded in Twitter tweets to predict the entire stock market is promising (Bollen et al., 2010; Zhang et al., 2011), very few established works (except for (Vu, Chang, Ha, & Collier, 2012)) provide concrete evidence to support that a single stock price can be correlated with Twitter sentiments. Lacking of such evidence may be due to the following reasons: (1) extracting sentiment features from the Twitter tweets is non-trivial. Tweets are short blog posts written by users. The length of a tweet is up to 140 characters. Thus the information is not explicitly expressed, and sometimes hidden in the URL links associated with the tweets. (2) Tweets relevant to an individual stock are not abundant enough for prediction tasks. Twitter users and sophisticated stock buyers are usually not the same group of people. Our untested conjecture is that Twitter users are more likely to be interested in technology stocks, since Twitter itself is representative of new technology trends. Nevertheless, even so, in our own study, we find only Apple Inc. stocks are often discussed in tweets; all other stock is not as often discussed (see Section 7.1). (3) Tweets related to a single stock may not about the company but its product. For example, tweets that mention “Facebook” are probably not suitable for stock price prediction because they may refer to the product Facebook that people use every day instead of the company. Instead, in our study, we used the dollar symbol “$” followed by stock symbols such as “FB” as a key word for searching stock related tweets of Facebook.

Due to these difficulties, researches on correlating and predicting stock prices with the Twitter sentiments embedded in the tweets that mention the stock are less likely to be fruitful. In fact, many prior works avoided the limited number of sentiment tweets by applying sentiment analysis on all public tweets data (not even related to stock market) and performed the Twitter sentiment analysis in a simplified way. For instance, Zhang et al. (2011) searches sentimental words such as “hope”, “happy”, “fear” and “worry” to determine the public mood in the tweets. As such, the correlation between individual stock prices and sentiments in the relevant Twitter data are largely overlooked. As a result, very little help can be offered in existing approaches to the stock investors when it comes to predicting an individual stock price.

Our study strives to explore to what extent these difficulties can be addressed. In other words, we want to apply sentiment analysis on the tweets mentioning or relevant to a particular stock and see how much sentiment information we can extract from the tweets being posted during an hour (e.g., 10:00am to 11:00am in a trading day). Next, we want to convert the sentiment data into structured Twitter features, which when combined with other metadata features could be correlated with the price changes of this underlying stock. We set our task to predict whether a particular company’s stock price will go up or down at the end of each hour, given all tweets collected during this hour. The best system will be the ones with the highest prediction accuracy. Through the task, we want to be able to answer the following questions: (1) which Twitter features are correlated with stock price and trading volume; (2) whether we can use the selected Twitter features to predict stock price directions.

3. Background

In this chapter, we review the background theories and prior work.

3.1 Stock Market and Efficient Market Hypothesis

In finance, the efficient-market hypothesis (EMH) asserts that the financial market is “informationally efficient”. “Weak”, “semi-strong”, and “strong” are the three major forms of the hypothesis. The weak form EMH claims that the current price already embedded all “past” information and thus analyzing past prices cannot predict future prices. The semi strong form of EMH claims that the current prices rapidly reflect all publicly available information and thus excess returns cannot be earned by fundamental analysis. In strong-form efficiency, current prices reflect all public and private information and no one can earn excess returns.

3.2 Related Work

3.2.1 Stock market and public mood

The correlation between human mood and the movement of stock market has been studied for decades. Variables, such as weather, length of daylight, lunar phases and temperature, have been considered to have impacts on human mood and therefore have been correlated with stock market in previous literature. Saunders (1993) conducted early studies on the influence of investor psychology, affected by local weather in New York

City, on stock prices. Similar positive effects of good weather on human mood was later confirmed and extended by Hirshleifer and Shumway (2003). The length of daylight has been recognized as another important factor of human mood, and study by Kamstra, Kramer, and Levi (2003) pointed out that seasonal affective disorder is correlated with the seasonal cycle of stock returns. Zheng, Yuan, and Zhu (2001) conducted a study on the effects of lunar phases on the stock market in 48 countries and concluded that stock returns are 3% to 5% lower on the days around a full moon than on the days around a new moon. Temperature is considered by Cao and Wei (2005). It has been studied in the psychology community that lower temperature is correlated to risk-taking behavior. Their study evidences that lower temperature leads to higher stock returns and thereby confirms the relation between human mood and stock prices. Edmans, Garcia, and Norli (2007) argued that a mood variable could be used to rationalize stock returns only when it is powerful enough to affect a large portion of investors. In their study, they calculated returns on the national stock market index during the first trading day after four types of major international sport matches (cricket, rugby, ice hockey and basketball), and found the returns are 38 point lower in average if the country losses the game. It further ruled out the effects of other factors such as loss of revenues and reduction in productivity on the stock market and confirmed that the movement in the stock market is purely due to public sentiment. However, the study found no correlation between wins of the games and the stock price movements. In our project, we focus on the emotion change of those Twitter users who tweet about certain stocks. Our conjecture is that the users tweeting about a certain stock are likely to invest in the stock as well. Therefore, the sentiment they expressed in their tweets can be used to predict the future stock prices.

3.2.2 Twitter and stock market

Using twitter data to predict stock market prices is an emerging topic. One reasonable rational behind the approaches is the relation between public mood and Twitter tweets (Bollen, Mao, & Pepe, 2010). As such, the hypothesis is that public mood expressed in Twitter tweets can be used to predict movements of the entire stock market. Sprenger and Welpe (2010) presented their work-in-progress study in which sentiment of tweets is associated with stock returns and volume of messages is associated with trading volume. Bollen et al. (2010) studied the correlation between public mood expressed in twitter tweets and Dow Jones Industrial Average. Instead of collecting tweets for a particular company or stock, the study makes use of all tweets that contain "I feel" or "I am feeling" or things alike to determine public moods. Two text mining tools are used in this research: OpinionFinder and GPOMS, which employs text mining techniques to determine, from the tweets data, positive or negative attitude, or six different mood (Calm, Alert, Sure, Vital, Kind and Happy) respectively. Granger causality analysis is used to find out the correlation between public mood and DJIA over time. The results indicate that "Calm" is most indicative of predicting DJIA, and it works better in combination with "Happy". Surprisingly, in their study, positive or negative sentiment is not directly correlated with DJIA.

Similarly, Zhang et al. (2011) randomly sampled one hundredth of all tweets during six months and measured the aggregated emotion. They found that the percentage of emotion tweets (both positive and negative) negatively correlate with stock market indicators such as Dow Jones, NASDAQ and S&P 500, but positively correlate with Chicago Board Options Exchange Market Volatility Index. However, the paper simply

uses emotional words such as “hope”, “happy”, “fear” and “worry” to indicate emotion within a tweet. Such approach oversimplifies the sentiment analysis of twitter data. We will use more sophisticated approach for sentiment analysis.

While the tweets-mood-stock models proposed by Bollen et al. (2010) and Zhang et al. (2011) are promising, when it comes to predicting individual stock prices, a few other features of twitter data have been analyzed to determine individual stock price changes. Ruiz et al. (2012) extracted features of twitter activities and used them to correlate with stock price and traded volume. The authors took a graph-based approach, in which the active tweets, users, hashtags and URLs in a day were connected as nodes in a graph. Edges in such graphs represent for relationships of nodes such as “annotate”, “retweet”, “mention”, “cite” and “create”. Then different features can be generated from the graph. The most indicative feature for trade volume as shown in this study is the number of connected components and the number of daily tweets. These two features also slightly correlate with the daily closing price. Most features used in this study are quantitative features, such as number of tweets in a day. Yi (2009) presents a research in the Master’s thesis demonstrating correlation between daily closing value of a stock and twitter data, represented in various models, e.g. frequency counting, loose n-gram models and noun phrase expansion. A more recent study by Mao et al. (2012) simply correlates daily number of tweets that mentions S&P 500 with S&P 500 closing price and achieves positive results. They also found the daily number of tweets that mention Apple Inc.'s stock strongly correlated with the trade volume and absolute price change.

3.2.3 Other media and stock prices

Orthogonal to our study is using information content in media types other than Twitter to predict stock market prices. Although these researches are not directly applicable to twitter data, the underlying concept is similar to ours. One such media in question is news articles. Lavrenko et al. proposed a language model to represent patterns of language that are correlated with stock behavior and then identify news stories related to the company that are indicative of stock trends (Lavrenko, Schmill, Lawrie, & Ogilvie, 2000). Pessimism about stock market in Wall Street Journal articles is used to predict movements of market prices by Tetlock (2007). High pessimism, according to the study, of the media will be followed by a downward in stock price and reversion to the fundamentals thereafter. Unusual pessimism, either high or low, can be correlated with high trading volumes. Hayo and Kutan (2004) reported positive correlation between energy news and stock returns in Russian financial markets, but no correlation between news and stock market volatility. Schumaker and Chen (2009) explored a predictive machine learning method for financial news articles analysis, which helps estimate a discrete stock price twenty minutes after a news article was released. They compared several textual representations of financial news articles and proposed a Support Vector Machine based approach to stock price prediction. They concluded that combining content in financial news and current stock price results in the best prediction performance.

Another well-studied media is financial message board and online stock discussion forum. Wysocki (1998) presented his findings in correlating message-posting volume about 3000 stocks in Yahoo! discussion boards with stock market activities.

Instead of demonstrating prediction power, the paper discussed relations between posting volume and short-term stock trading behavior changes. Tumarkin and Whitelaw (2001) correlated activities in online stock discussion forum, ragingbull.com, with stock prices of a few Internet service companies. The results of the study were in support of the theory of market efficiency in that the message-board activities couldn't predict the stock price in the following day. Using Internet message-board activity to predict stock market was also studied by Antweiler and Frank (2004). By analyzing 1.5 million messages posted on Yahoo! Finance and Raging Bull about 45 companies in DJIA and Dow Jones Internet Index, the paper concluded that stock messages only helped predict stock volatility; the prediction power on stock returns is economically small.

Weblogs, or blogs, are yet another type of sources of information that can be derived to predict stock market. Choudhury, Sundaram, and Seligmann (2010) studied that communication dynamics in the blogosphere, e.g. number of posts, number of comments and etc., and correlates them with stock market movement. Gilbert and Karahalios (2010) presented a study in which emotion estimated from weblogs can be used to predict stock market prices. The study estimates the anxiety, worry and fear from 20 million weblogs on LiveJournal and concludes that the widespread worry could be negatively correlated with S&P 500 index.

3.2.4 Evaluation methodology

Various techniques have been used in previous works to evaluate the effectiveness of the proposed approaches. Some takes correlation-only approaches in which the major purposes of these studies were finding the correlation between the features they selected and the stock prices (Bollen et al., 2010; Choudhury et al., 2010;

Gilbert & Karahalios, 2010; Zhang et al., 2011). Some uses more sophisticated statistical analysis. For instance, Bollen et al. (2010) calculated mean absolute percentage error as evaluation method, and Yi (2009) used the simple moving average. Nevertheless, more commonly used approach to evaluate the effectiveness of using text-mining approach to predict stock prices are direct prediction accuracy and investment return simulation. More specially, prediction accuracy in terms of price change directions was used as evaluation methods (Bollen et al., 2010; Mao et al., 2012; Schumaker & Chen, 2009). Simulation based evaluation approach, in which an automated investor is modeled to buy or sell stocks based on the proposed algorithm, was adopted (Lavrenko et al., 2000; Mao et al., 2012; Schumaker & Chen, 2009). In this thesis, we used prediction accuracy as the metric to evaluate the performance of prediction. The prediction accuracy specifies the percent of predictions in which the tasks of classifying Twitter features associated with positive stock price movements and negative stock price movements are correct. Hence if the prediction accuracy is higher than the baseline, which is the percentage of the majority class in the testing dataset, we conclude that the prediction is more powerful than a “naïve” predictor which simply guesses the majority class every time.

4. Twitter and Tweets Metadata

Micro-blog is a new type of social media, which has shown a potential in facilitating information exchange. A micro-blog is essentially a stream of short messages that is written by a single user and shared among large amount of readers. Current popular micro-blogging services include Facebook, Tumblr, Twitter, etc. Because of Twitter's widespread use, it has become the most popular micro-blogging platform nowadays.

4.1 Twitter

Twitter, created in 2006, is an online micro-blogging service, which allows users to post and share their own text-based message in less than 140 characters each time. The user can get access to the service via many ways, such as the Twitter.com website, mobile application, and etc.

One distinctive feature of this micro-blogging platform is the real-time updating and widely reaching mechanisms. Because of its capability of releasing news information rapidly, Twitter has been used for a lot of purposes in a variety of scenarios. For example, it has been used to organize protests, such as the 2009 Iranian presidential election protests, 2011 Egyptian revolution, and etc. Twitter is also used as an effective de facto emergency communication system for breaking news.

Another feature of Twitter is the relationship between users. The follow-and-be-followed relationship allows user to subscribe to each other and get their up-to-date

updates rapidly. In such ways, news can be passed along from one user to another and broadcasted to more readers in very short time.

4.2 Tweets Metadata

On the Twitter platform, a user can post tweets, follow other users and be followed by other users. She can also create lists to include other users so that any status change of these users will be seen immediately. Accordingly, the user can also be listed by other users. A tweet post by a user can be original and retweet of other user's tweets. A tweet can contain hashtags, the "#" symbol, which is used to mark keywords or topics in a Tweet. Similarly, Twitter users are recommended to use "$" symbol before stock symbols when mentioning stocks.

These functionalities require each tweet to contain metadata. Actually, Twitter data contains more information than the tweet itself. Each tweet can be much larger in size than 140 characters. It also contains the metadata, specifying the statistics information about the tweets. The metadata contains information about both the tweet and the user who posted the tweet. For instance, when using streaming API, a tweet is comprised of, but not limited to, the following metadata:

Table 1

Twitter Metadata

Metadata	Meaning
created_at	The time at which the tweet was created by the user
uid	A string of numbers specifying the unique ID of a tweet.
text	The tweet message itself
source	The client software from which the user posted the tweet, web, smartphone, or somewhere else.
truncated	whether the length of the tweet has been truncated

	due to character limits
entities	The URL, user name, or hashtag included in the tweet
in_reply_to_status_id	The ID string of the tweet that this tweet is replying to
in_reply_to_user_id	The user ID string that the tweet is replying to
name	The name of the user who posts the tweet
user_created_at	The time at which the user account was created on Twitter
followers_count	how many followers that the user has
friends_count	how many users that the user is following
listed_count	how many lists the user is included
statuses_count	how many tweets have been posted by the user since the account was created

In this thesis, we collected tweets about 15 technology stocks in NASDAQ. Mishne and Rijke (2006) said people are inclined to engage more in technology and political related information on social media. Therefore, we believe choosing technology companies to be our predicting targets will help us obtain sufficient relevant tweets data.

5. Experiment Design

In this chapter, we will introduce the experiment design in our study. More specially, we will first discuss our data collection process. Then we will sketch our data processing process, which includes stock selection and feature selection. Finally, we will discuss how we are going to evaluate our approaches.

5.1 Data Collection

In order to study the correlation between stock prices and Twitter data, two sets of data were collected during our experiment: stock trading prices and Twitter tweets. The sources of the data are described as below.

5.1.1 Stock Prices

Stock price data can be collected from many sources, such as Google Finance and Yahoo! Finance. However, only daily open prices, close prices, highest/lowest prices and daily trade volume are available for free for historical stock prices. In order to obtain finer grained stock prices, one needs to collect data in real time. As of this writing, only Yahoo! Finance still provides API for automated data collection. The stock prices provided by Yahoo! are updated every 5 minutes. We thus developed an automated stock price checker, which uses Yahoo! Finance API to retrieve stock prices every 5 minutes from 9:30 to 16:30 eastern time during March 12, 2013 to June 6, 2013. The collected data includes the time stamp, stock name, the current stock price and trading volume.

We selected 15 technology stocks in NASDAQ. In Table 2, we listed stock symbols, company names and market cap.

Table 2
Stock Symbols and Market Capital

Stock Symbols	Company Names	Mkt Cap (billion)
AAPL	Apple Inc.	424.53
AMD	Advanced Micro Devices, Inc.	2.88
CSCO	Cisco Systems, Inc.	130.24
CTXS	Citrix Systems, Inc.	12.18
FB	Facebook Inc.	59.36
GOOG	Google Inc.	288.89
INTC	Intel Corporation	120.35
LNKD	LinkedIn Corp	18.69
MSFT	Microsoft Corporation	292.54
NTAP	NetApp Inc.	13.70
NVDA	NVIDIA Corporation	9.02
ORCL	Oracle Corporation	161.76
SNDK	SanDisk Corporation	14.57
VMW	VMware, Inc.	30.63
ZNGA	Zynga Inc.	2.72

5.1.2 Tweets

In terms of tweets, there are two ways to collect data in our research. One is to collect the tweets stream directly from twitter.com in real time, and then use the collected data for research. However, we cannot use this approach to study stock behavior in the past, because since July 2011, Twitter has changed its historical tweets access policy. Even search of historical tweets for academic research purpose is not allowed any more. An alternative approach is to get historical data from non-Twitter sites, such as Datasift, Gnip and Topsy. It also seems possible to use Google search for tweets.

For this study, we set up a server, which connects to Twitter.com via streaming API. The Twitter streaming API returns public tweets that match the specified filter predicates. More than one keyword is allowed so that only a single connection is required

for data collection. The key words matching algorithm in the streaming API is case insensitive. That is, searching for “Twitter” will return results containing “twitter” or “TWITTER”. In addition, special characters before or after the searched key words will also be included. For example, searching for “Twitter” may get results containing “#twitter” or “$twitter”. Twitter recommend users use “$” symbol together with the stock symbol when mentioning stock prices. However, user may simply use the stock symbol regardless the recommendation.

Using company name as key words in our study may include unrelated tweets. For example, use “Apple” as key words can get results like “I like eat apples”. Using “google” will get tweets referring to products of the Google Inc. And other scholars also used the dollar symbol to retrieve stock (Ruiz et al., 2012; Yu & Kak, 2012). We also found that only use stock symbols in searches may get tweets not specific to the stock prices. As such, filtering collected data becomes very difficult. As such, we use the dollar symbol “$” followed by stock symbols such as “AAPL” as a key word for searching stock related tweets of Apple Inc. Similarly, we search '$FB' for tweets related to the Facebook stocks. The downloaded tweets will also contain Twitter metadata (see Section 5.2.2) together with the text of the tweets. All tweets related data were stored in MySQL database for future references.

5.2 Data Processing

During the data processing, we first extracted tweets from the MySQL database and then aggregated the tweets for OpinionFinder to process. Then we created a separate table in MySQL database to store the number of positive words and negative words in each tweets. Each tweet can be correlated with the previous table using the str_id field of

the tweets. The next step is to perform correlation analysis for feature selection. The selected features were then used for stock price prediction.

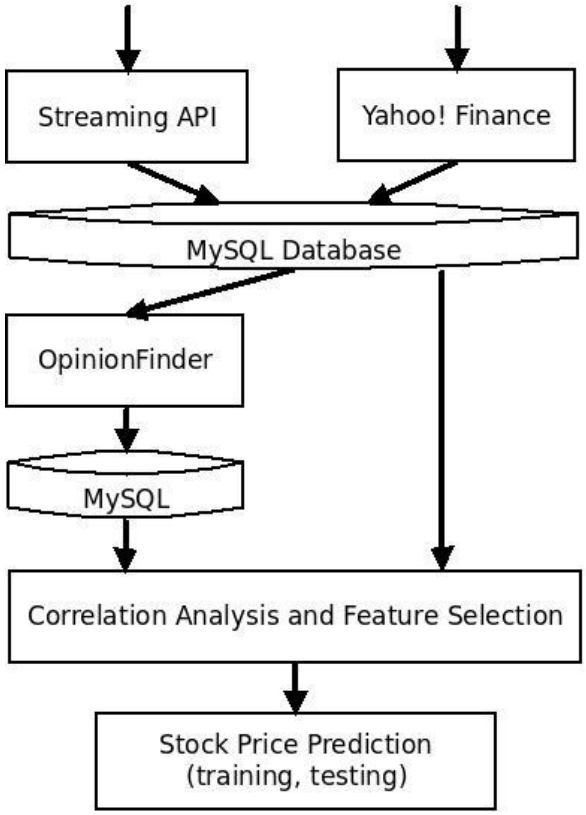

Figure 1. Data processing work flow.

5.2.1 Sentiment analysis

OpinionFinder takes a list of documents as inputs to process. In particular, each document contains the content of exactly one tweet. The outputs are sets of files, in the format of SGML/XML markup language. Each file contains results related to one aspects of the input document. We were particularly interested in the file, exprclass.polarity, which reports the occurrence of positive words and negative words respectively. We used $W_{positive}$ and $W_{negative}$ to denote the number of positive words and negative words in each tweet respectively and insert the results into a separate MySQL table named

"sentiment". The key field of table "sentiment" is the "str_id" of the tweet that can be used to correlate with the metadata of the same tweets.

After tweets collection and the sentiment analysis, we get Table 3 that shows the number of the tweets, and the ones with positive or negative sentiments in descending order for these 15 technology companies. All tweets listed in Table 3 are collected between 10:00am to 16:00pm every trading day.

Table 3
Stock Symbols, Number of Tweets, and the Ones with Sentiments

Stock Symbols	N_{tweets}	$N_{positive}$+$N_{negative}$
AAPL	42625	4622
GOOG	12291	1149
FB	8454	838
MSFT	5847	397
LNKD	2704	307
INTC	2123	157
ZNGA	1369	140
CSCO	1188	157
ORCL	1099	113
AMD	926	89
VMW	773	80
NVDA	559	32
SNDK	537	33
NTAP	318	11
CTXS	146	18

5.2.2 Feature extraction

During feature extraction, we tried to aggregate Twitter statistics of the tweets that were generated during each hour. Specially, we focused on the New York Stock Exchange operating hours from 10:00 to 16:00 eastern time from Monday to Friday, excluding holidays. Accordingly we collected tweets from the time period and separate them into 6 hours: 10:00-11:00, 11:00-12:00, 12:00-13:00, 13:00-14:00, 14:00-15:00, 15:00-16:00. The data processing is accomplished with programs written by us, which

extract data from MySQL databases and aggregate relevant features and then summarize the data in .csv format that can be recognized by SPSS for correlation, or Weka for classification.

5.2.3 Correlation analysis

Not all Twitter features are strong indicators of future stock prices. Therefore, we first use SPSS Statistics to analyze the correlation between each feature and stock prices or trading volume. We used Pearson correlation coefficient to indicate correlation relationship, with two-tailed test of significance. The Pearson correlation coefficient is a measure of the linear correlation between two variables, returning a value between +1 and −1. The larger the coefficient is, the better the two variables are correlated; the smaller the level of significance is, the more confident the correlation results are. The correlation coefficient is calculated by:

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

5.2.4 Prediction

With selected Twitter features, we used Weka to perform classification of the two classes: positive price change and negative price change. We used logistic regression classifier.

In statistics, logistic regression is usually used for predicting the outcome of a categorical dependent variable based on one or more independent variables. Though logistic regression can be binomial or multinomial, it is usually used to refer specifically to the instance in which the observed outcome is binary—that is, the available categories

have only two possible types. In our case, the outcome is coded as “up/positive” and “down/negative”. Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables that are usually continuous, which are tweet metadata in our thesis.

5.3 Evaluation Methodology

For each stock we investigated, we only make predictions when we observe sentiment tweets. Therefore, given the sentiment tweets we collected during each hour, we try to make predictions on the stock price by the end of that hour. Because the data we used in our study is unbalanced, we find accuracy of the prediction must be compared with the baseline accuracy, which is the percentage of the majority class in the testing dataset. Because the most simply classifier would be guessing the majority class every time which, though is meaningless, still achieves a prediction accuracy higher than 50%.

Using prediction accuracy as the metric indicates the percent of predictions that successfully foresee the stock price movement direction, rather than the amount of changes. We consider this approach because the amount of stock price rises or drops, we suspect, not only correlates with public opinion on the stock but also related to the previous stock price and even the stock market as a whole. As a result, in our study, we only try to correlate with price change directions and as long as our classifier yields higher prediction accuracy than the baseline accuracy, we can conclude the Twitter sentiments combined with Twitter metadata may have positive correlation and certain prediction power with the stock markets.

6. Feature Selection

In previous research works of predicting stock prices using Twitter data, two major types of features were collected and used to correlate stock market. They are sentiment features and metadata features. In this chapter we will discuss how we select these two types of features.

6.1 Sentiment Features

The Twitter users may have strong opinion or sentiment when editing their tweets. Such sentiments, if successfully extracted and analyzed, can be very useful tools to study user's attitude toward certain events, products and stocks. As shown in prior studies, such sentiments can reflect general believes of the Twitter users and potentially affect the stock market. Some prior works made use of Twitter sentiments implicitly. For example, Yi (2009) explored the use of bag-of-word model for stock price correlation. The rationale behind his/her model is that certain words express the user's opinion and mood more than other words and thus have high probability to indicate the future movements in the stock market.

Sentiment features are unstructured. A tweet may or may not contain sentiments in its content. Even a tweet with obvious sentiment bias may be hard to recognize correctly. In our study, we used existing text mining software to extract sentiment features. We analyzed the sentiment features of tweets using OpinionFinder (Wilson et al., 2005), which is open source software that uses a pipeline of tools to perform

subjectivity analysis. The text-based data-mining task is conducted by automatically identifying opinion sentiments and speculations in text. The tool conceptually split the text-mining task into two parts, document processing and sentiment analysis. Document processing is performed with OpenNLP to tokenize and parse sentences and with SCOL for stemming. SUNDANCE is used to identify patterns for extraction. In the sentiment analysis phase, WordNet is used as a subjective expression and speech event classifier and a Naïve Bayes classifier for subjective sentence is built based on BoosTexter machine learning program.

Given the text of a tweet, we used OpinionFinder to determine the occurrence of sentimental words in the tweet. The output specifies the locations of the sentimental words, if any, and whether it is positive sentiments or negative sentiments. We aggregated the sentimental words and report the number of positive sentiments and negative sentiments respectively. As such, unstructured sentiment features are converted into structured features.

6.2 Metadata Features

The metadata features used in our study were from the tweet statistics and Twitter metadata, which are aggregated from the sets of tweets predicted positive or negative respectively. From all the Twitter metadata we analyzed, we selected to use the following features to further study their correlation with stock prices.

***Number of tweets*: N_{tweets}**

N_{tweets} specifies the number of tweets that we used streaming API to collect during each hour in which the stock symbol was mentioned. More N_{tweets} means more mentions among people and more attentions to this stock.

Number of tweets with sentiments: $N_{positive}$ *and* $N_{negative}$

Not every tweet contains sentiment features that can be recognized by our sentiment analysis tool. $N_{positive}$ and $N_{negative}$ specify the number of tweets with positive sentiment and negative sentiment respectively. Therefore, usually $N_{tweets} \gg N_{positive} + N_{negative}$. We believe the numbers of sentiment tweets are important because sentiment tweets expressed users' opinion on the stock, thus may have influence on other user's future buying or selling behavior.

Number of Followers: $N_{pos\_follower}$ *and* $N_{neg\_follower}$

$N_{pos\_follower}$ (or $N_{neg\_follower}$) is the sum of the follower numbers of the users who posted positive (or negative) tweets mentioning the underlying stock during an hour The more followers a user has, the more influence the user may have through a single tweet. Therefore, the greater the value $N_{pos\_follower}$ has, the more Twitter users are potentially influenced by the positive mood shared by the tweets during this period. Similarly, the greater the value $N_{neg\_follower}$ has, the more Twitter users are potentially influenced by the negative sentiments.

Number of Friends: $N_{pos\_following}$ *and* $N_{neg\_following}$

$N_{pos\_following}$ (or $N_{neg\_following}$) is the sum of the friends numbers of the users who posted positive (or negative) tweets mentioning the underlying stock during an hour. A friend is another Twitter user that a user is following in Twitter. Our untested hypothesis is that Twitter user that has large number of friends can be influenced by other users.

Listed count: N_{pos_listed} *and* N_{neg_listed}

N_{pos_listed} (or N_{neg_listed}) is the sum of user created lists that the users who posted positive (or negative) tweets mentioning the underlying stock during an hour are included in. The more lists the user is included, potentially the more influence the user may have through a single tweet. We expect these features to have similar effects as $N_{pos\_follower}$ (or $N_{neg\_follower}$).

Status count: N_{pos_status} *and* N_{neg_status}

N_{pos_status} (or N_{neg_status}) is the sum of total statuses update (tweets) that the users, who posted positive (or negative) tweets mentioning the underlying stock during an hour, have posted since their accounts were created. The more tweets a user has posted, the more likely their tweets will be seen and paid attention to by their followers.

User history: $N_{pos\_user\_history}$ *and* $N_{neg\_user\_history}$

$N_{pos\_user\_history}$ (or $N_{neg\_user\_history}$) is the sum of the numbers of days that the user accounts, who posted positive (or negative) tweets mentioning the underlying stock during an hour, have been created. The longer a user account is created, the more trust potentially their followers may have in their tweets.

User activities: $N_{pos\_user\_activity}$ *and* $N_{neg\_user\_activity}$

$N_{pos\_user\_activity}$ (or $N_{neg\_user\_activity}$) is the sum of the average tweets per day posted by the users, who posted positive (or negative) tweets mentioning the underlying stock during an hour. This feature is created to compensate $N_{pos\_user\_history}$ (or $N_{neg\_user\_history}$) and N_{pos_status} (or N_{neg_status}), because users with longer history may post fewer tweets.

7. Experiment Evaluation

The Twitter data and stock prices used in our experiment evaluation were collected from Mar 12, 2013 to June 6, 2013.

7.1 Correlation

In this section, we evaluate how tweets sentiment correlates with stock price changes. Particularly, we analyzed 5 most popular stocks: AAPL, MSFT, GOOG, LNKD, FB, because according to Table 3, these 5 stocks have comparably greater number of total tweets and number of tweets that have sentiments. Therefore, we believe their tweets could provide more information about the stock price, which may make the correlation more reliable. We separated the data related to each stock into two files, recording the tweet statistics related to positive and negative price changes, respectively.

We show the correlation results between the features and the positive stock price change in Table 4, the features and the trading volume during positive stock change in Table 5. Those features related to negative stock price changes are illustrated in Table 6 and the correlation between features and the trading volume during negative stock change are shown in Table 7.

From Table 4 below, we can see that the number of tweets (N_{tweets}) is strongly correlated with the positive price changes. The number of positive sentiment tweets is also correlated with price changes. Since the price change is positive, the correlation

between negative sentiment tweet number ($N_{negative}$) and the price is much weaker. However, quite unexpectedly, the number of followers ($N_{pos\_follower}$, $N_{neg\_follower}$) and number of lists the user is in, positive or negative (N_{pos_listed}, N_{neg_listed}), are not correlated with the stock price. What is really surprising is that the positive user history ($N_{pos\_user\_history}$) and positive user status (N_{pos_status}) are both strongly correlated with the positive stock changes.

Table 4
Correlation of Twitter Features and Positive Stock Price Change

Feature Price	AAPL		MSFT		GOOG		LNKD		FB	
	R	p	R	p	R	p	R	p	R	p
N_{tweets}	.424	.000	.278	.007	.217	.069	.524	.000	.369	.001
$N_{positive}$	.275	.018	.184	.079	.205	.086	.360	.002	.395	.001
$N_{pos\_follower}$	.011	.927	-.023	.827	-.015	.900	.288	.013	.198	.095
N_{pos_listed}	.064	.588	.043	.684	-.008	.945	.270	.021	.198	.096
N_{pos_status}	.206	.080	.049	.643	.142	.239	.393	.001	.402	.000
$N_{pos\_user\_history}$	.358	.002	.118	.261	.157	.190	.351	.002	.352	.002
$N_{pos\_user\_activity}$	.040	.738	.045	.668	-.002	.990	.412	.000	.245	.038
$N_{negative}$	.228	.052	.083	.429	.103	.393	.361	.002	.313	.007
$N_{neg\_follower}$	-.024	.839	.007	.945	-.092	.445	.144	.226	.248	.036
N_{neg_listed}	.016	.896	.031	.768	-.074	.540	.189	.109	.299	.011
N_{neg_status}	.112	.347	.052	.625	.199	.095	.285	.014	.310	.008
$N_{neg\_user\_history}$	.202	.086	.085	.418	.162	.176	.254	.030	.308	.009
$N_{neg\_user\_activity}$	.026	.830	-.073	.489	.022	.857	.280	.016	-.024	.841

From Table 5 below, we can see that N_{tweets}, $N_{positive}$, N_{pos_status}, $N_{pos\_user\_history}$, $N_{negative}$, N_{neg_status}, and $N_{neg\_user\_history}$ are all strongly correlated with trading volume when the price change direction is positive. This means when the stock is going up, both positive tweets features and negative tweets features are correlated with the volume.

Table 5
Correlation of Twitter Features and Volume of Positive Stock Price Change

Feature Volume	AAPL		MSFT		GOOG		LNKD		FB	
	R	p	R	p	R	p	R	p	R	p

N_{tweets}	.599	.000	.355	.001	.569	.000	.741	.000	.697	.000
$N_{positive}$	.602	.000	.123	.244	.385	.001	.485	.000	.604	.000
$N_{pos\_follower}$	.072	.548	-.032	.765	-.092	.447	.155	.190	.279	.018
N_{pos_listed}	.178	.132	.083	.434	-.082	.495	.224	.057	.259	.028
N_{pos_status}	.464	.000	.062	.554	.156	.193	.318	.006	.545	.000
$N_{pos\_user\_history}$	.653	.000	.076	.474	.303	.010	.435	.000	.586	.000
$N_{pos\_user\_activity}$	.027	.818	.055	.599	.065	.589	.220	.062	.329	.005
$N_{negative}$	.561	.000	.216	.038	.440	.000	.524	.000	.614	.000
$N_{neg\_follower}$	.110	.356	-.017	.874	-.096	.426	.146	.219	.340	.003
N_{neg_listed}	.209	.076	.000	1.0	-.053	.658	.157	.184	.433	.000
N_{neg_status}	.506	.000	.078	.463	.516	.000	.092	.439	.608	.000
$N_{neg\_user\_history}$	.591	.000	.234	.025	.405	.000	.461	.000	.622	.000
$N_{neg\_user\_activity}$	.108	.363	.030	.774	.293	.013	.052	.662	.055	.647

Table 6 and 7 are shown below to illustrate the correlation between the features and the price change or volume during negative price change.

Table 6

Correlation of Twitter Features and Negative Stock Price Change

Feature Price	AAPL		MSFT		GOOG		LNKD		FB	
	R	p	R	p	R	p	R	p	R	p
N_{tweets}	-.582	.000	-.307	.020	-.204	.064	.008	.944	-.056	.621
$N_{positive}$	-.573	.000	.082	.546	.035	.757	-.017	.884	.039	.733
$N_{pos\_follower}$	.008	.947	.116	.390	-.103	.355	.091	.426	.078	.493
N_{pos_listed}	.026	.821	.111	.411	-.030	.785	.093	.414	.056	.624
N_{pos_status}	-.364	.001	-.045	.737	.063	.573	.071	.537	-.050	.661
$N_{pos\_user\_history}$	-.549	.000	.100	.461	-.006	.957	.005	.964	-.017	.880
$N_{pos\_user\_activity}$	-.091	.420	-.186	.167	.002	.989	.046	.689	-.117	.880
$N_{negative}$	-.594	.000	-.159	.239	-.289	.008	.010	.931	.038	.740
$N_{neg\_follower}$	-.288	.009	-.289	.029	-.206	.061	.062	.588	.093	.411
N_{neg_listed}	-.332	.002	-.233	.082	-.181	.102	.097	.394	.021	.852
N_{neg_status}	-.490	.000	-.089	.508	-.268	.014	.058	.610	-.132	.243
$N_{neg\_user\_history}$	-.590	.000	-.093	.489	-.299	.006	.042	.716	-.010	.929
$N_{neg\_user\_activity}$	.085	.449	-.042	.754	-.074	.504	.049	.669	-.279	.012

Table 7

Correlation of Twitter Features and Volume of Negative Stock Price Change

Feature Volume	AAPL		MSFT		GOOG		LNKD		FB	
	R	p	R	p	R	p	R	p	R	p
N_{tweets}	.630	.000	.660	.000	.198	.073	.267	.017	.618	.000
$N_{positive}$	.637	.000	.114	.398	.208	.059	.130	.253	.181	.107
$N_{pos\_follower}$	-.087	.439	-.038	.780	.165	.136	-.147	.196	-.100	.379
N_{pos_listed}	-.100	.372	-.078	.565	.170	.124	-.155	.174	-.084	.459

N_{pos_status}	.482	.000	.104	.443	.149	.180	-.134	.238	.164	.147
$N_{pos\_user\_history}$	.612	.000	.082	.545	.179	.106	-.010	.927	.186	.098
$N_{pos\_user\_activity}$	-.037	.743	.227	.090	.082	.462	-.119	.296	.194	.085
$N_{negative}$	.713	.000	.613	.000	.232	.035	-.076	.508	.109	.337
$N_{neg\_follower}$	.313	.004	.582	.000	.157	.157	-.044	.701	-.028	.804
N_{neg_listed}	.345	.002	.629	.000	.194	.078	-.020	.859	.081	.476
N_{neg_status}	.522	.000	.614	.000	.213	.053	-.017	.880	.253	.024
$N_{neg\_user\_history}$	.697	.000	.472	.000	.226	.040	-.068	.550	.252	.024
$N_{neg\_user\_activity}$	.108	.363	.434	.001	.032	.771	-.076	.507	.208	.064

Except for AAPL in Table 6 or AAPL and MSFT in Table 7, where the correlation between features such as N_{tweets}, $N_{positive}$, N_{pos_status}, $N_{pos\_user\_history}$, $N_{negative}$, N_{neg_status}, $N_{neg\_user\_history}$, and negative price changes are still obvious, it is hard to find correlation in other stocks. This can be explained by looking at the total number of tweets and total number of sentimental tweets, shown in Figure 2 and Figure 3. In Figure 2, the total number of tweets mentioning AAPL is much larger than any other stocks, which makes correlating metadata features with AAPL stock price more accurate.

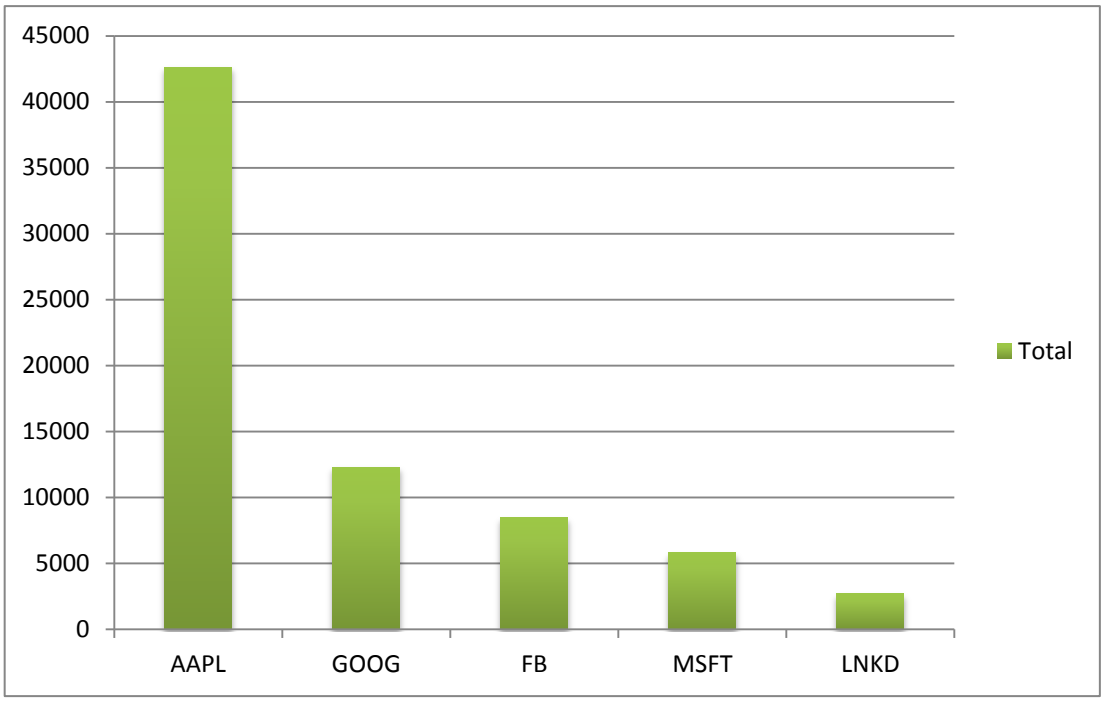

Figure 2. Total number of tweets.

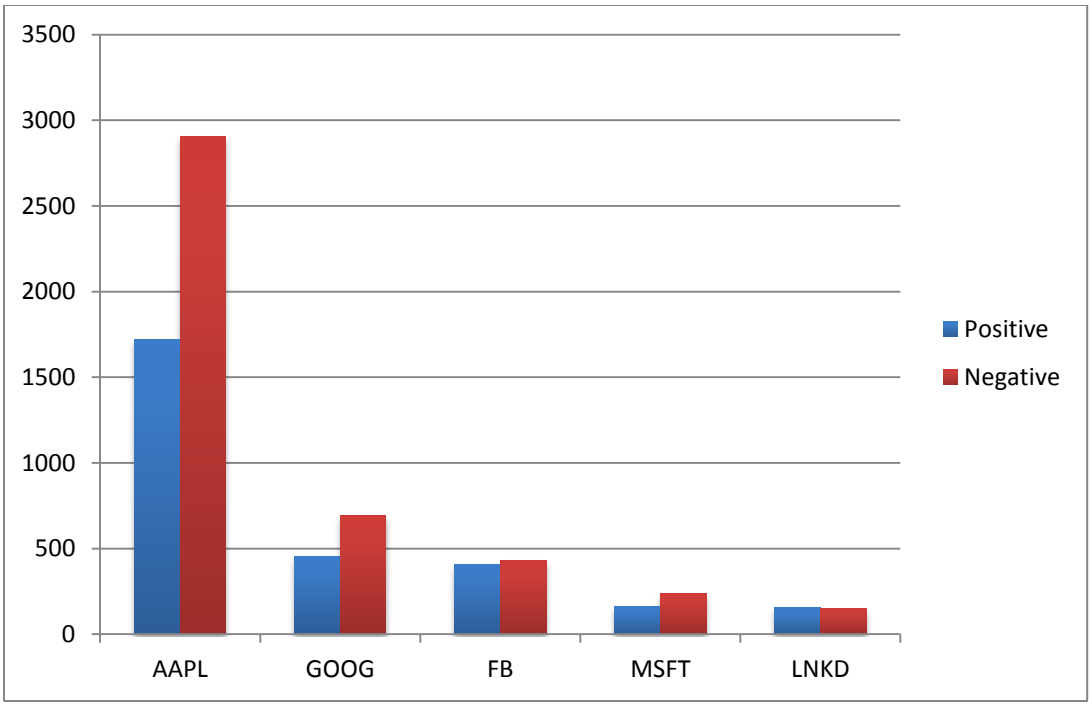

Figure 3. Number of tweets with sentiments.

In sum, we chose to select N_{tweets}, $N_{positive}$, N_{pos_status}, $N_{pos\_user\_history}$, $N_{negative}$, N_{neg_status}, $N_{neg\_user\_history}$, as features to predict stock prices.

7.2 Prediction

Next we use the features selected in Section 7.2 to predict stock prices by the end of each hour. Specially, we selected the hours during Mar. 12 to Jun. 6 in which $N_{positive}$ and $N_{negative}$ are not zero at the same time, and then collected the features selected above to correlate with the stock price change by the end of each hour on the hour compared to the stock price on the last hour. The prediction task is essentially a classification process: given the features extracted from Twitter data, will the stock price goes up or down?

7.2.1 Evaluating Prediction Results.

Using the logistic regression classifier, we first performed 10-fold cross validations on the 15 stocks collected from Mar. 12 to May 12. On each day, tweets from 10:00 to 16:00 eastern time were aggregated using the method described as in Section 6.2. The data for each stock is slightly imbalanced. We define the baseline of the dataset as the percentage of the majority class in the dataset. It is because a "dummy" classifier can simply predict the majority class in each prediction than render accuracy higher than 50%. The baseline in the data we collected for this evaluation ranges from around 50% to 60%. In Table 8, the prediction accuracy reported by Weka was provided.

Table 8
Prediction Accuracy

Stock	Accuracy	Baseline
AAPL	0.4883	0.5352
AMD	0.5604	0.5934
CSCO	0.5144	0.5048
CTXS	0.4730	0.5135
FB	0.5117	0.5587
GOOG	0.5305	0.5164
INTC	0.5613	0.5896
LNKD	0.5429	0.519
MSFT	0.6056	0.6244
NTAP	0.5036	0.5108
NVDA	0.5509	0.5868
ORCL	0.5707	0.5288
SNDK	0.5494	0.5432
VMW	0.5523	0.5465
ZNGA	0.5423	0.5473

We can see from Table 8 that for over half (nine) of the stocks, our prediction does not outrun the baseline accuracy. In fact, the average prediction accuracy for the 15 stocks was 53.7% while the average baseline for the 15 stocks was 54.8%. This means

the predictions are not successful in those stocks, although the prediction accuracy exceeds 50% for most of the stocks.

7.2.2 Feature Filtering

The next step in our experiments was to determine which feature provides the most information about the stock prices. Thus we evaluated experimentally how such prediction results change when using fewer features. For that purpose, we used the same data set as in Section 7.2.1 and removed exactly one feature from the dataset for classification each time and then run 10-fold cross validation on the data. The corresponding prediction accuracies are listed in the Table 9.

Table 9
Prediction Accuracies with One Less Feature

Stock	N_{tweets}	$N_{positive}$	N_{pos_status}	$N_{pos\_user\_history}$	$N_{negative}$	N_{neg_status}	$N_{neg\_user\_history}$	Baseline
AAPL	0.512	0.502	0.531	0.516	0.474	0.531	0.469	0.535
AMD	0.549	0.588	0.588	0.577	0.544	0.560	0.527	0.593
CSCO	0.481	0.505	0.505	0.505	0.519	0.514	0.534	0.505
CTXS	0.459	0.473	0.459	0.473	0.446	0.500	0.446	0.514
FB	0.512	0.540	0.502	0.516	0.521	0.521	0.531	0.559
GOOG	0.512	0.488	0.526	0.498	0.493	0.479	0.502	0.516
INTC	0.566	0.571	0.575	0.575	0.580	0.575	0.580	0.590
LNKD	0.557	0.524	0.533	0.538	0.552	0.538	0.543	0.519
MSFT	0.610	0.592	0.610	0.596	0.620	0.606	0.615	0.624
NTAP	0.482	0.504	0.504	0.504	0.504	0.504	0.518	0.511
NVDA	0.581	0.551	0.557	0.551	0.545	0.551	0.545	0.587
ORCL	0.578	0.539	0.565	0.534	0.581	0.571	0.565	0.529
SNDK	0.549	0.543	0.543	0.543	0.549	0.549	0.562	0.543
VMW	0.564	0.558	0.547	0.552	0.529	0.552	0.558	0.547
ZNGA	0.502	0.572	0.552	0.547	0.532	0.537	0.542	0.547
Avg.	0.534	0.537	0.540	0.535	0.533	0.539	0.536	0.548

In Table 9, the last row indicates the average prediction accuracy when removing the corresponding feature from the dataset for classification. Roughly speaking, removing these features does not change the prediction accuracy much. This means none of the

features is really indicative to the stock price changes. However, we do found that removing features of N_{pos_status} and N_{neg_status} will result in slightly better prediction accuracy. This may suggest that the total number of tweets that a user has posted so far has nothing, or very little, to do with the influence of this user's most recent sentimental tweets. Therefore, we eliminate these features from further experiments.

7.2.3 Evaluating with Testing Datasets

The experiments in Section 7.2.3 used cross validation as an evaluation method. In this section, we want to further experiment with a separate testing dataset.

Therefore, in the following experiment, we first train the logistic regression classifier with the two-month Twitter data (Mar. 12 to May 12, 2013) and then test it against Twitter data and stock prices collected from May 13 to Jun. 6, 2013. The prediction accuracies for the 15 stocks in such settings and their corresponding baselines are listed in Table 10.

Table 10
Prediction Accuracies and the Baselines in the Testing Dataset

Stock	Accuracy	Baseline
AAPL*	0.545	0.586
AMD	0.571	0.643
CSCO	0.486	0.600
CTXS	0.750	0.750
FB*	0.618	0.640
GOOG*	0.526	0.505
INTC	0.313	0.531
LNKD*	0.462	0.558
MSFT*	0.541	0.581
NTAP	0.364	0.545
NVDA	0.333	0.500
ORCL	0.5	0.800
SNDK	0.75	0.750
VMW	0.4	0.600
ZNGA	0.529	0.588

From Table 10, we can see that the prediction accuracy is even worse than cross-validation. Even the five stocks that have most tweets (marked with * in Table 10) don't get accuracy higher than baseline. The results are expected. In a cross validation, the training data and testing data has similar baselines, that is the ratio of positive cases and negative cases are roughly the same in training and testing. However, with a different testing dataset, the ratio may change drastically, and the prediction accuracy may be affected.

7.2.4 Discussion

From our experiments, we found that our selected features, even though positively correlates with stock prices, do not successfully predict stock price changes. The failure in prediction may due to many reasons. It may be caused by an inaccurate model we built to represent Twitter sentiments, or insufficient data in our study. While the Twitter sentiment model we used in our paper is the basic hypothesis we intend to test here, insufficient experiment data may be a possible cause. As tweets mentioning a specific stock is very rare, the sentimental ones are fewer. It is an insurmountable problem for predicting individual stock price using the proposed method. Therefore, the unsuccessfulness in prediction may either caused by lack of sufficient data, or it proves that our method to extract Twitter sentiments and to combine with Twitter metadata is problematic, which means our hypothesis that using the approach proposed in this paper to combine Twitter sentiment data and Twitter metadata can predict stock price changes may not be correct.

Nevertheless, it may also suggest the impossibility of applying Twitter sentiment analysis for stock prediction because there is no causal relation between the two.

This may also mean that using Twitter data in stock prediction over-simplifies the problem. Other factors need to be considered, such as the history stock prices or the entire stock market movements. In fact, there may be too many factors that can affect the stock market which is not reflected by the Twitter data.

8. Conclusion

In this paper, we explored combining Twitter sentiment data with its associated metadata to correlate with individual technology stock prices, and making predictions on the future price changes with selected Twitter features. More particularly, we used OpinionFinder, an existing text-mining technique, to extract Twitter sentiment data from plaintext tweets and then aggregate Twitter metadata associated with sentimental tweets. The results are features extracted from Twitter data that indicate the amount of Twitter sentiments relevant to a specific stock during a given period of time (an hour). We next use the features to correlate with stock prices at the end of the each period. In our evaluation, we found that features such as N_{tweets}, $N_{positive}$, N_{pos_status}, $N_{pos\_user\_history}$, $N_{negative}$, N_{neg_status}, $N_{neg\_user\_history}$ are positively correlated with stock price changes and trading volumes. The features that are positively correlated with the stock prices are selected to make predictions, using machine learning algorithms, on the future stock price movements. Our results of the prediction, however, are not as successful as expected. Most of the classification results were not even as good as the baseline, which a "dummy" classifier can easily achieve by simply guessing the majority class. The unsuccessfulness of our evaluation suggests that our hypothesis that Twitter sentiments can be used to predict individual stock prices may be wrong. But in order to prove it, we still have quite a few other possibilities to rule out.

Using Twitter data to predict stock market is still an ongoing research. Much of

the results are too early to be used for real stock market trading strategies. Whether Twitter sentiments can be used to predict stock price is still a hypothesis that is yet to be tested. Our results just provide one piece of negative evidence to such hypothesis. In practice, modern trading institutes still employs more sophisticated prediction models, which also consider the historical stock prices and macroeconomics. It is likely that Twitter data is more suited to improve stock market prediction, rather than making decisions on its own. Future work is required to further explore different possibilities and make stronger conclusions.

Reference

Antweiler, W. & Frank, M. Z. (2004). Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. *The Journal of Finance, 59(3),* 1259-1294.

Bollen, J., Mao, H., & Zeng, X. (2010). Twitter mood predicts the stock market. *Journal of Computational Science, 2(1),* 1–8.

Bollen, J., Mao, H., & Pepe, A. (2010). Determining the public mood state by analysis of microblogging posts. *Proceedings of the Proc. of the Alife XII Conference,* Odense, Denmark, MIT Press.

Cao, M., & Wei, J. (2005). Stock market returns: A note on temperature anomaly. *Journal of Banking & Finance, Elsevier, 29(6),* 1559-1573.

Choudhury, M.D., Sundaram, H., John, A., & Seligmann, D.D. (2010). Can Blog Communication Dynamics be Correlated with Stock Market Activity? *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia,* 55-60.

Edmans, A., Garcia, D., & Norli, O. (2007). Sports sentiment and stock returns. *The Journal of Finance, 62(4),* 1967–1998.

Fama, E. (1965), The Behavior of Stock Market Prices. *Journal of Business, 38,* 34–105.

Gilbert, E., & Karahalios, K. (2010). Widespread worry and the stock market. *In Proceedings of the International Conference on Weblogs and Social Media*.

Hayo, B., & Kutan, A. M. (2004). The Impact of News, Oil Prices, and Global Market Developments on Russian Financial Markets. William Davidson Institute Working Papers Series, 656, William Davidson Institute at the University of Michigan.

Hirshleifer, D. A., & Shumway, T. (2003). Good Day Sunshine: Stock Returns and the Weather. *The Journal of Finance, 58(3),* 1009–1032.

Kamstra, M., Kramer, L., & Levi, M. (2003). Winter blues: A SAD stock market cycle. *American Economic Review, 2003,* 324-343.

Lavrenko, V., Schmill, M., Lawrie, D., & Ogilvie, P. (2000). Mining of Concurrent Text and Time Series. *In Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* Workshop on Text Mining, 37–44.

Malkiel, B. G. (1973). *A Random Walk Down Wall Street* (6th ed.), W.W. Norton &

Company, Inc.

Mao, Y., Wang, B.,Wei, W., & Liu, B. (2012). Correlating S&P 500 Stocks with Twitter Data. HotSocial12, Beijing, China.

Mishne, G., & Rijke, de M. (2006). A Study of Blog Search, *In Proceedings of 28th European Conference on Information Retrieval (ECIR)*.

Ruiz, E. J., Hristidis, V., Castillo, C., Gionis, A., Jaimes, A. (2012). Correlating Financial Time Series with Micro-Blogging Activity. *In Proceedings of WSDM12,* Seattle, Washington, USA.

Saunders, E. M., Jr. (1993). Stock Prices and Wall Street Weather. *The American Economic Review, 83(5),* 1337-1345.

Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions on Information Systems (TOIS), 27(2).*

Schumaker, R. P., & Chen, H. (2009). Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFinText System. *ACM Transaction Information Systems, 27,* 1-19.

Sprenger, T. O., & Welpe, I. M. (2010). Tweets and trades – the information content of stock microblogs. *Social Science Research Network Working Paper Series*.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance, 62,* 1139–1168.

Tumarkin, R., & Whitelaw, R. R. (2001). News or noise? Internet message board activity and stock prices. *Financial Analysts Journal, 57,* 41–51.

Vu, T. T., Chang, S., Ha, Q. T., & Collier, N. (2012). An Experiment in Integrating Sentiment Features for Tech Stock Prediction in Twitter. *In Proceedings of the Workshop on Information Extraction and Entity Analytics on Social Media Data*, 23–38.

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., … Patwardhan, S. (2005). OpinionFinder: A system for subjectivity analysis. *In Proceeding of HLT/EMNLP on Interactive Demonstrations*, 34-35.

Wysocki, P. D. (1998). Cheap Talk on the Web: The Determinants of Postings on Stock Message Boards. University of Michigan Business School Working Paper No. 98025.

Yi, A. (2009). *Stock Market Prediction Based on Public Attentions: a Social Web Mining Approach.* Master's thesis, University of Edinburgh.

Yu, S., & Kak, S. (2012). A survey of Prediction Using Social Media. *ArXiv e-prints*.

Zhang, X., Fuehres, H., & Gloor, P. A. (2011). Predicting Stock Market Indicators Through Twitter "I hope it is not as bad as I fear". *Procedia - Social and Behavioral Sciences, 26(0),* 55 - 62.

Zheng, L., Yuan, K. Z., & Zhu, Q. (2001). Are Investors Moonstruck? - Lunar Phases and Stock Returns. Working papers series.