Predicting Equity Returns Using Twitter Sentiment

by

Mitchell Dean Muhlheim

(Under the Direction of Dr. Cheolwoo Park)

Abstract

Predicting events and responding to unforeseen events quickly has been a significant struggle. Previous publications have shown the ability to monitor and predict events using semantic analysis. These publications have been able to use semantic analysis to predict movements in the stock market and respond to events faster than ever before. This thesis aims to fuse these ideas to develop and test a new approach to semantic scoring and test its effectiveness. The semantic scoring focuses on mood related word referencing a single company, a sector, and the market for predicting stock returns. In addition to returns, focus was placed on predicting direction of the market as well. This work found an inconclusive relationship between sentiment and the returns over the following days. With the framework from this thesis in place, a refined wordlist and modeling could improve predictive accuracy.

Index words: stock market prediction, sentiment analysis, twitter, mood, equity

Predicting Equity Returns Using Twitter Sentiment

by

Mitchell Dean Muhlheim

B.S., Berry College, 2009

A Thesis Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Master of Science

Athens, Georgia

2013

Predicting Equity Returns Using Twitter Sentiment

by

Mitchell Dean Muhlheim

Approved:

Major Professor:	Dr. Cheolwoo Park
Committee:	Dr. Kang Li
	Dr. Lily Wang

Electronic Version Approved:

Dr. Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2013

Predicting Equity Returns Using Twitter Sentiment

Mitchell Dean Muhlheim

April 26, 2013

Acknowledgments

I owe my deepest gratitude to my parents for instilling in me a passion for knowledge and education, and without whom I would not be the person I am today. To my sister for supporting my every endeavor. To Whitney for her boundless love and compassion. To Andrew for the countless hours of water-cooler talk which led to the formulation of this thesis.

I would also like to thank my committee for continuously pushing me to develop the ideas contained herein. And to my other friends and family who helped make the completion of this project possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Predicting the stock market is a new and exciting endeavor reaching through many fields. With applications in statistics, computer science, linguistics, finance, and psychology, researchers are trying to build models capable of equity prediction. With the advent of web based applications such as Twitter and Facebook, new and extremely large datasets are accessible for researchers to examine. These Web 2.0 tools along with forums and message boards have changed how investors receive information and subsequently, how they invest. Groups of people can invest irrationally based on mood, but with instantaneous dissemination of information, there exists a larger potential pool of participants to act in accordance with mood. While acting with mood instead of analysis, market participants will purchase stocks in line with sentiment instead. Many participants acting in unison with similar feelings on a security can influence its price.

Population sentiment is broadcast every second from market participants and nonparticipants alike. Analysis of population sentiment can be accomplished by tapping into the Web 2.0 data stream. With a sufficiently fast algorithm for sentiment scoring, one can tap into a real-time feed of user mood. And with a pulse on user mood, predicting buying patterns is the next logical step.

The goal of this thesis is to discover a new sentiment scoring methodology and its application to stock market prediction. This thesis proposes a new multiplicative methodology for sentiment scoring with special attention paid to processing time. This new methodology is more robust to current vernacular and allows a finer scale by which sentiment per sentence is judged. With the new methodology, models testing predictive capability of sentiment analysis related to stocks are explored. Prior models used multiple regression with little consideration to temporal relationships between observations. This model attempts to account for influence as well as time by using a time series regression. In Chapter 2, a review of background and current literature are discussed. Chapter 3 covers the creation of a new method for mood scoring and how data were collected. Chapter 4 addresses model creation and validation and Chapter 5 concludes and suggests future work.

Chapter 2

Literature Review

The social media revolution has allowed people share their thoughts, ideas, and feelings, anywhere in the world, in seconds. Petabytes of information are freely shared by users every year, 140 bytes at a time [1]. It is the ability to freely and instantaneously access this information that is of interest. It can be seen through recent research that a fundamental shift in predictive modeling is on the horizon. Through exponential increases in modern computing power and further study into semantic analysis, there is a very real possibility of better predicting human behavior. Many different approaches and methodologies exist hoping to extract the most useful and relevant information from users. Several papers have been written recently chronicling the power that can be harnessed from Twitter and the importance of the information discovered. Twitter can be used to respond to events in almost real time, successfully gauge population sentiment, and predict the future.

2.1 Finance Background

Eugene Fama famously applied a random walk model to the stock market in the 1960s. This application was called the Efficient Market Hypothesis (EMH) [2]. Because news is random

and unpredictable, the market will follow a random walk pattern and is thus unpredictable beyond 50 percent accuracy. Similarly, because the market is efficient, all publicly available information is instantaneously absorbed into the price. There are large organizations with faster connections than any individual user, and they are capable of exploiting mispriced securities and arbitrage opportunities through automated trading. Thus, any new information or model is automatically accounted for in the pricing due to changes in demand. Another cornerstone of the EMH is that it relies on the rationality of market participants to act in their own best interest. Not until the 1990s did behavioral economists begin gaining traction in disproving the fundamental assumptions of EMH. The first is the disproval of the random walk theory [3]. It has been shown that the market does not follow a naive random walk model, and in fact is predictable about 65 percent of the time using machine learning algorithms. Advances in processing power have allowed for more complex models as well as utilizing multiple high volume data streams. Bollen et al. (2011) showed that Twitter user sentiment is capable of predicting the directional movement of the DOW Jones Industrial Average, used as a proxy for the market, 87 percent of the time [4]. The second is the evidence that market participants do not always act rationally [5]. Individuals tend to purchase past winners and hold onto past losers in bullish markets, while selling off both in bearish markets. Participants are also subjected to both positive and negative external influence from myriad sources such as newspapers, blogs, micro-blogs, and television. These influences lead to higher volume of purchasing in light of good news, and higher volume of sales due to bad. Investors tend to overreact to news and market signals, and in turn, drive prices too far in one direction [6]. The fusion of irrational investors and inefficient markets leads to a new field of predictive modeling of equity markets. Not only investigation of the stock market, but of all seemingly random events. Terrorist attacks [7], stock market returns [8], and box office receipts [9] are all being explored through predictive modeling relying on semantic analysis.

2.2 Reliability of Semantic Analysis

In the quickly expanding field of semantic analysis, it can be seen that there is some useful insights are lurking within the seemingly infinite stream of words. It started with simplistic lexicons and returning positive sentiment or negative sentiment on a plus one, zero, minus one scale. The corpus, or large set of text, was coded by hand for sentiment scoring. This method was expensive, time consuming, and very tedious. The corpus also utilized a small volume of tweets, typically in the hundreds or thousands. New access to Twitters API, along with increased computing speeds, and a better understanding of internet vernacular, has shifted data collection, analysis, and results to new heights. Understanding how Twitter users speak is an important step to deciphering their mood. Individual mood can impact decision making, creativity and memory. Using semantic analysis on tweet text has been shown to significantly track mood of individuals over time [10]. The plus one, zero, minus one scale method for matches in a lexicon, is capable of detecting positive or negative sentiment quite reliably. Mood detection from text has taken cues from psychologists in lexicon and development. In [4], the authors were able to apply a program called Google-Profile of Mood States (GPOMS) to their Twitter stream and successfully captured a six-dimensional profile of the user base. GPOMS works with a lexicon matching method, which captures six different mood states of calm, alert, sure, vital, kind, and happy. In bench-marking GPOMS, they were able to see a significant decrease in calm moods and a significant increase in vital moods the day before the 2008 Presidential Election. The day of the elections, those moods returned to their normal state, and a significant increase in happiness and surety was observed. The day after the election, happiness returned to its normalized baseline until Thanksgiving, when a significant spike occurred. This new methodology appears to accurately gauge user sentiment for the population of English speaking Twitter users.

2.3 Predicting Events Using Twitter

Because of the real-time nature of Twitter, it is possible to estimate mood almost instantly. In [7], researchers were able to apply sentiment analysis to Twitter users in the early responses to the Mumbai terrorist attacks. They found through their framework, they could chronicle user sentiment and track response in near real-time. The importance of this result is exciting. Events, as they happen, are discusses almost instantly on Twitter. Changes in mood state are almost instantly recognizable with software monitoring the stream of tweets. Similarly, in [11], the authors show the ability for users to disseminate information faster than currently relied upon channels. In the aftermath of an earthquake, the United States Geological Survey has shown that Twitter users reports can actually track location of seismic activity faster than traditional seismographs. Furthermore, crisis management response and organization can be established within minutes instead of hours. While faster than its traditional counterparts, predicting events via web channels is also exceptionally accurate. Researchers found that web search queries can provide information long before trusted channels. Google Inc. and the Centers for Disease Control and Prevention (CDC) have shown that users search queries coupled with location data can be used to uncover and track flu epidemics with great accuracy [16]. Individuals will typically only search for flu like symptoms when they are exhibiting those symptoms. Thus, the simple count of searches based on the flu is adequate in tracking an epidemic. Again, similar to earthquakes, these data are available long before hospitals can synthesize and report the patient intake data. Semantic analysis can instantly estimate mood and mood is linked to predicting behavior, so is it possible to use mood estimates to predict behavior in near real time? A white paper from Hewitt Packards Social Computing Lab shows evidence that not only is it possible, but it may be even better than other market predictors [9]. The authors logged average tweet rate about movies, finding a fairly high correlation between box office receipts and tweet rate. Further, they added the ratio of

positive tweets to negative tweets to the rate in a similar time series. They found, with extremely high correlation, a relationship between Twitter discussion and opening weekend box office receipts using this methodology.

2.4 Synthesis of Ideas

In several recent articles, the above ideas have been stitched together to create one very interesting idea. Many tools have been developed to capture tweets and score them based on semantic analysis in attempts to predict directional movement of the stock market. These articles use proxies for the market such as the Standard and Poor 500 (S&P 500) or the Dow Jones Industrial Average (DJI). The first step in all methodologies is to connect to Twitters Application Programming Interface (API). This allows unfettered access into Twitter. Once accessed, the tweets are cleaned by removing stop words [12]. These words make up a majority of English sentence structure, yet provide no actionable text for semantic analysis. The size of the lexicon of stop words varies from article to article, but can encompass up to 60 percent of text volume on Twitter. The cleaning involves removal of these words and will significantly reduce processing time for the cleaned text corpus. In the case of [4], researchers used GPOMS and OpinionFinder (OF) software to successfully predict Dow Jones movement with accuracy of 87 percent. By normalizing the scores, they created a baseline sentiment for general opinion and the six psychological mood states referenced earlier. The normalized scores allowed for easy computational and visual detection when opinion or mood state spiked. A multiple linear regression is then used to indicate significant coefficients. Those significant coefficients are then used in a time series to determine direction. In [8], the authors used the Twitter API to successfully predict stock market direction. The lexicon used in this approach involved ascribing bullish and bearish sentiment to tweets. The tweets were sourced from a website called StockTwits covering 3,874 authors. They were then aggregated

based on an index proposed in [13] called the Bullishness Index. Several models utilizing various mixes of price, volume, Bullishness Index, 5-day lagged price, 5-day lagged volume, and 5-day lagged Bullishness Index were used to predict upward or downward movement.

In conclusion, irrational market participants, the ability to successfully quantify mood from text, and the real-time nature of Twitter give researchers newfound predictive power into the stock market. The availability of instantaneous response shows great potential for research. Developing a model which records sentiment and monitors events while reacting in real time is a very real possibility for researchers.

Chapter 3

Methodology

This research was performed in order to examine a link between equity price returns and the sentiment about the company to which the stock belongs. Data were collected by gathering a sample of tweets from Twitter utilizing a specific query involving keywords pertaining to the company. This event happened every minute, so it was imperative that analysis was completed in less than a minute. The combination of several methods was employed to achieve this desired result.

3.1 Data Capture

Data were captured using Twitters API by a PHP script that parsed the site for a specific query. The query involved keywords relating to the company such as the name, ticker symbol, and main products the company produced. Only English results were included in order to maximize the reliability of semantic analysis. The PHP script was executed every minute and returned 100 of the most recent tweets. Due to the limitations of the Twitter query framework, a single IP address is only allowed 200 queries per hour. Furthermore, a maximum of 100 tweets can be returned per page, unless a further loop function is added for

page 2 and beyond. At 200 queries an hour, a query could conceivably be run about once every 20 seconds. For most queries with relatively low volume, this would surely cause some overlap of tweets.

The results were split into observation number, time recorded, tweet text, and tweet id number, and then stored in a MySQL database. The MySQL table can be exported to a CSV file for further analysis in R or SAS. PHP and MySQL were chosen because of the low cost of implementation, capability to act as a bridge between the web and statistical analysis programs, and ability to run on multiple machines. An example of a search for apple would return these sample results:

1. Five years ago today, Apple launched the iPhone and began the era of everyone touching their cell phones http://cnet.co/MdHyzs

2. Apple tweaks iOS App Store search algorithm again with return to keywords, names

3. Tom Cruise and Katie Holmes to split. However, Apple assures customers that their daughter Suri will continue to answer your questions.

3.2 Data Cleaning

A downside to text based analysis is that it is computationally expensive. An increase in text quantity greatly increases processing time, so any methods to reduce overall text were considered. The constraint for this analysis was to have the tweets fully analyzed within a minute, in order for the system to be ready for the next incoming batch a minute later. A benefit to using MySQL was the ability to clean the data prior to importing it into a statistical analysis package. For example, some tweet text contained tabs, escapes, and characters which R could not read. Before the data were read from the database, a cleaning function was run in order to remove characters which inhibited further analysis. Further

data cleaning was implemented through the removal of stop-words. These words are those which add to the sentence length, but do not impact the meaning of the sentence. For example, and, that, this, and so are all stop-words which add a significant amount of text in the English language, but do not influence sentiment. All of the stop-words were removed prior to sentiment analysis, which resulted in cutting processing time in half. Using the example test and cleaning the data, the following result is produced:

1. years ago Apple launched iPhone began era everyone touching cell phones http://cnet.co/MdHyzs
2. Apple tweaks iOS App Store search algorithm return keywords names
3. Tom Cruise Katie Holmes split however Apple assures customers daughter Suri continue answer questions

3.3 Creating a Tweet String

Another method to reduce processing time was to collapse the 100 tweets into one large string of text [12]. Since the employed method will rely on dictionary based text matching, it was faster to match occurrences within one long string instead of 100 short strings. This approach simply pasted the tweets end to end, where they were ready to be analyzed. The tweet string of the above example would be:
"years ago apple launched iphone began era everyone touching cell phones http://cnet.co/MdHyzs apple tweaks ios app store search algorithm return keywords names tom cruise katie holmes split however apple assures customers daughter suri continue answer questions".

3.4 Text Matching

The text matching portion of the analysis relied on two separate dictionaries with similar structures. The first dictionary was a list of 1200 commonly used adjectives in the English

language [20], and scores representing the degree of positivity or negativity for the corresponding word by using a scale from negative 2 to positive 2. This approach attempts to enhance commonly employed plus minus methods by adding a finer scale by which a word can be scored. For example, the word 'best' has a higher degree of positive sentiment than the word 'good', and would be represented by a positive two, instead of a positive one. Consumers talking on a public forum about a product being the best will carry more weight with readers than if the product were simply 'good' or 'adequate'.

The second dictionary contained commonly used adverbs and, what for the remainder of this thesis will be called modifier words. This dictionary contained 250 words [22] that could enhance or reverse the meaning implied by the tweet. These words were also scored on a scale from negative two to positive two, but applied slightly differently. Upon a match of an adjective from the first dictionary, a match from the modifier dictionary would be sought. In the event of a successful match in both dictionaries, the scores for each word would be multiplied. Score multiplication helps reduce falsely matching positive words when intended to be negative. In negative sentiment case of 'not good', a nave approach would have only read good and assigned a positive one. With the multiplicative approach, the word 'not' is assigned a negative one, and the word 'good' is assigned a positive one. The multiplication of the modifier and adjective scores results in the desired sentiment score of negative one.

3.5 Method Capability

Testing the method against another medium was an important step to see if sentiment could actually be quantified by counting adjectives and multiplying them by a modifier. Amazon was the chosen test bed. In one sitting, users of Amazon write their sentiment about a product, and then score it from one to five stars. This appears to be a great test whereby the text can be scored via the methodology above, and then directly compared to the star

rating of the users own sentiment.

We applied the proposed methodology above to some selected Amazon reviews. Only reviews of three paragraphs of similar size and structure were used. In the test cases, an obvious difference between reviews was observed. The one star reviews attained slightly negative to slightly positive scores, while the five star reviews attained much higher scores. The range of sentiment scores was -8 to $+32$ for these test cases. The slightly negative lower bound can be explained by the relative lack of negative words compared to positive ones in the adjective and adverb dictionaries.

An example of one such review is the following 1 star reviews receiving a total score of 2:

"I owned my 4th gen ipod touch for 10 days before it's screen shattered. Did I drop it? NO! I was simply playing a favorite game on it one night while in bed. In the darkness I mistakenly pressed the rearward facing camera on the touch screen (mistaking it for the home button) the glass immediately disintegrated in a radial pattern centered around where I touched it. Really Apple? I mean I have been a faithful customer for years because your company produced the best quality swag even though it came at a premium. I called Apple to complain and they pretended that they haven't heard a thing about this. I told the person on the phone, "look buddy, google "broken 4th gen ipod touch screen", and then tell me you haven't heard anything about there being a problem. Here are the facts. My ipod "touch" broke because I was touching it. Apple will not refund my money, or replace the item. There is clearly a huge design/manufacturing flaw in the glass on this model. At this point I am hoping everyone else that this has happened to will file a class action suit against Apple. I mean, shees! way to rest on your laurels, Apple!"

And a sample of a similar size and structure 5 star review receiving a score of 10:

"The iPod Touch is such a wonderful device that i am having a hard time finding any flaws to it. Great screen, unbeatable resolution on such a tiny device. The device itself is very minimalist and it's super thin. AND it's BEAUTIFUL. New hardware revision made the bottom flat again, so there is no wobbling when you put it on a flat surface. I used to own an ipad and i was surprised to find that most apps that i bought are universal apps that also works with the ipod touch. Both the quality and quantity of the apps in the App Store are unbeatable. It's great for streaming video, i use Air Video to stream my video on my desktop to the ipod touch, and the app even converts non-natively supported formats like mkv on the fly. Overall, this is an amazing device that i will recommend to all my friends."

3.6 Creating a Total Score by Day

The timestamps for each block of 100 tweets were kept, and the resulting scores of each tweet summed over the minute long period. Using the chron package in R and keeping the timestamps with the minute scores, allowed for very easy summing of the data. Data were summed up in two ways. The first approach was to sum up the scores by the day. This resulted in the total daily score for the Twitter query which successfully captured the sentiment of 144,000 tweets.

3.7 Capturing General Mood

The first approach is to assess whether or not the proposed method is successful in predicting movement at the market level. A previously proposed method for capturing general mood queried Twitter for a very specific result in order to evaluate over all Twitter population mood [1]. In that approach, this query was successfully able to detect changes in mood on Twitter and predict movement in the Dow Jones Industrial Average. To test a change in sentiment scoring, the above methodology was applied to the same search parameters. The query searched for all tweets containing the text I am or I am feeling, and returns only those results. The proposed method is an expansion of that paper by using a larger dictionary and allowing for modifiers.

3.7.1 Capturing Market Price

The closing values were captured from Yahoo Finance for the Standard and Poor 500 Index (ticker: $\hat{\text{G}}$SPC). The S&P 500 is a collection of 500 global companies integral to the United States economy. It is generally regarded in finance as the best proxy for the market as a whole. In financial modeling, it is oftentimes referred to as the market for its close similarity to the entire stock market.

3.7.2 Capturing Sector Mood

Allowing for more specificity, it is of interest to examine the scoring methodology at the sector level as well. A more specific line of inquiry may show inconsistency in the model or validate is usage at even more micro levels. Capturing sector mood required the query to involve a sector which met two criteria. The first is a popular sector which is garnering attention of market participants. The second is that it is a trendy enough subject to warrant a high volume of tweets. For sector mood, Green Energy was decided to meet both of these

criteria. As of now, green energy is receiving a lot of press and is a very popular subject for social media discussion. This method used a query containing both green and energy, where both terms had to be present for inclusion.

To capture the sector pricing, the opening and closing values were captured from Yahoo Finance for an index called Wilderhill Clean Energy (ticker: PBW). This index is represented by the managing company as an accurate representation of the whole global alternative energy sector. It is balanced occasionally to keep its holdings in line with shifting global trends and appears to be a suitable proxy for capturing sector valuation. Due to the cutting edge nature of green technology, no generally accepted proxy for the sector exists.

3.7.3 Capturing Company Mood

Exploring this framework even more, it is of interest to see if this methodology can be applied at the company level. The query involved searching for the individual company name, major product lines, and the ticker symbol. Major product lines were chosen as they represent a significant contribution toward the mood of a company as well as the companys performance. The company chosen had to meet similar criteria to the sector mood criteria. The company must be publicly traded, have significant investment volume, and garner attention from individuals tech savvy enough to tweet about them. The company chosen was Apple, with the query terms apple, ipod, ipad, iphone, mac, macintosh, or AAPL. It is believed that these terms should adequately capture discussions on the company and its products, as well as generate significant volume of tweets for analysis.

3.7.4 Capturing Company Price

The closing values were captured from Yahoo Finance for Apples stock (ticker: AAPL). Apple is a very large company with an enormous market capitalization. It is also a very

technology focused organization with a user base generally interested in the bleeding edge of technology.

3.7.5 A Word on Capture

These terms produce a very high volume of tweets which well exceed the maximum capability of 100 tweets per minute of this framework. Thus, the tweets were captured by a de facto systematic sampling approach with $n = 100$ at every 1 minute interval.

Chapter 4

Analysis

The goal of this research is to establish a new methodology for sentiment scoring and discover any possible links to predictive power over the stock market. Due to incremental changes in scope, data were collected across time periods of varying size. First, a probing phase to test the hypothesis of mood normality was employed. This initial collection for company mood (Apple) was from January 13, 2012 to March 20, 2012 yielding $n = 5,314,848$ tweets. Next, the data collection for the general mood, and attached to the S&P 500, ranges from April 20, 2012 to September 20, 2012, yielding a total of $n = 14,118,339$ tweets. The data collection for sector mood (PBW) ranges from April 20, 2012 to September 20, 2012 yielding a total of $n = 13,719,116$ tweets. The data collection for company mood ranges from April 20, 2012 to September 20, 2012, yielding a total of $n = 14,314,848$ tweets. The scores are shown in Figures 4.1 and 4.2 and appear to be roughly stationary. The data, taken from Yahoo Finance (http://finance.yahoo.com/) from the same periods as above for their respective score, as described in Chapter 3. Figures 4.3 and 4.4 show the prices over the observed period. Because the prices are obviously non-stationary, they were then transformed by log-return, shown in Figures 4.5 and 4.6. The log return was used because it returns a stationary time series without losing important information. The formula for calculating log-returns is as

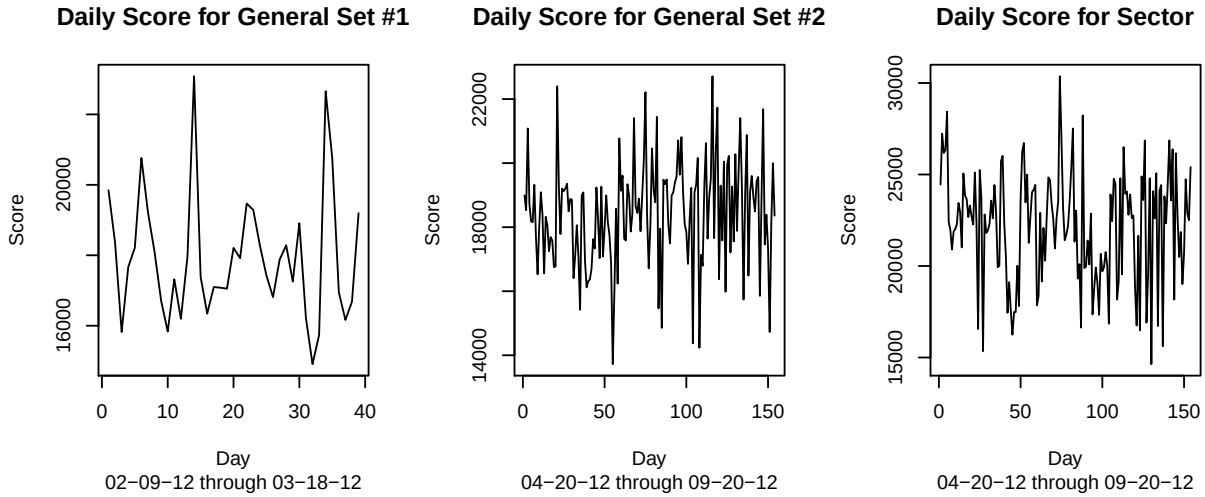

Figure 4.1: Daily Scores

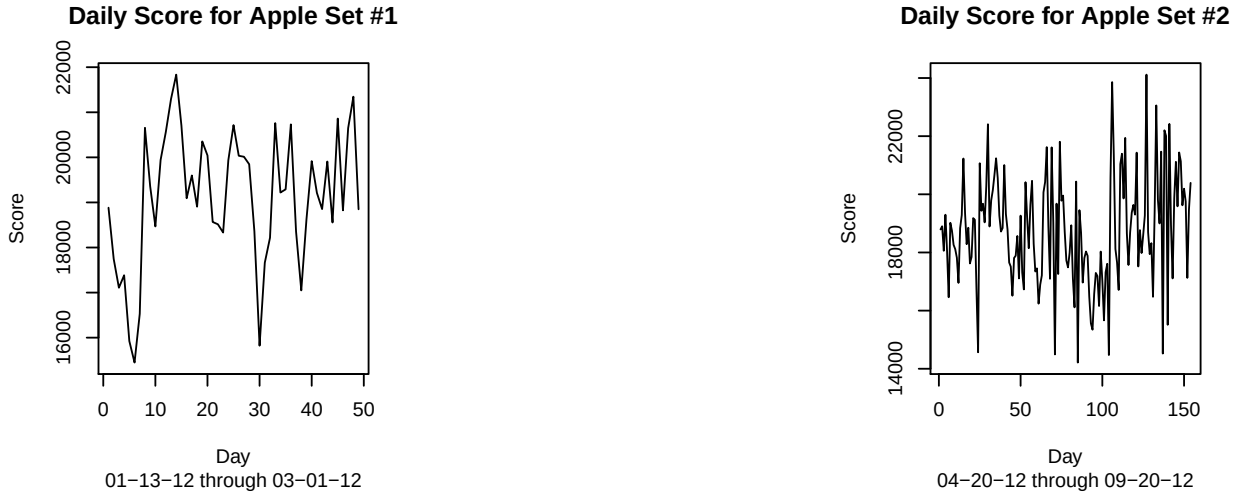

Figure 4.2: Daily Scores

follows:

$$R_t = \log(p_t) - \log(p_{t-1}) \tag{4.1}$$

where p_t is the price of the security or index at time t.

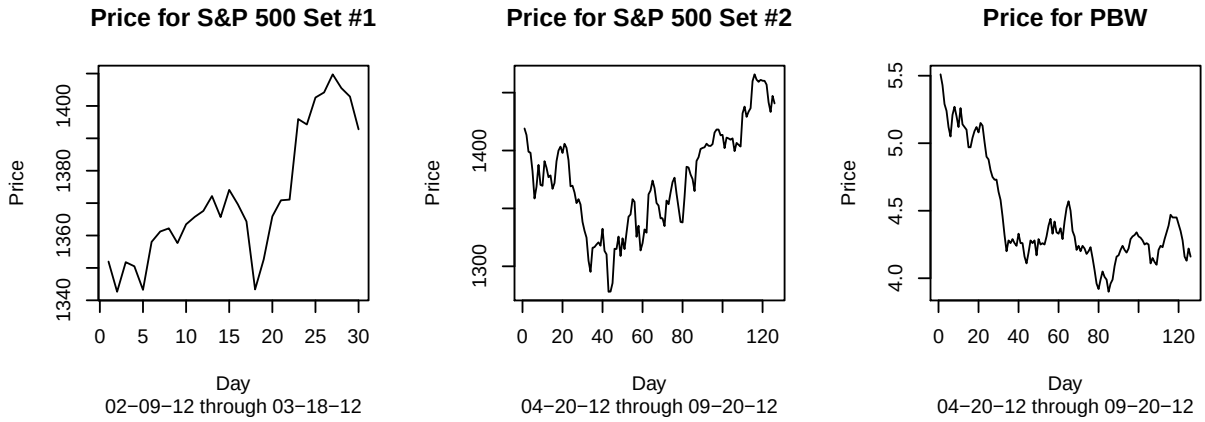

Figure 4.3: Daily Price Before Transform

Figure 4.4: Daily Price Before Transform

The first check of this proposed methodology was to see how the data are distributed. The general mood at any given day appears to be normally distributed. There would be observed significantly negative sentiment and significantly positive sentiment, but the majority of users would cluster around a central score. This is the case for all three of the observed sentiment scores. The histogram for the S&P500 Set #1 (Figure 4.7, left) is skewed slightly toward negative sentiment, with a few high sentiment spikes. This opposite is indicated by the 2nd

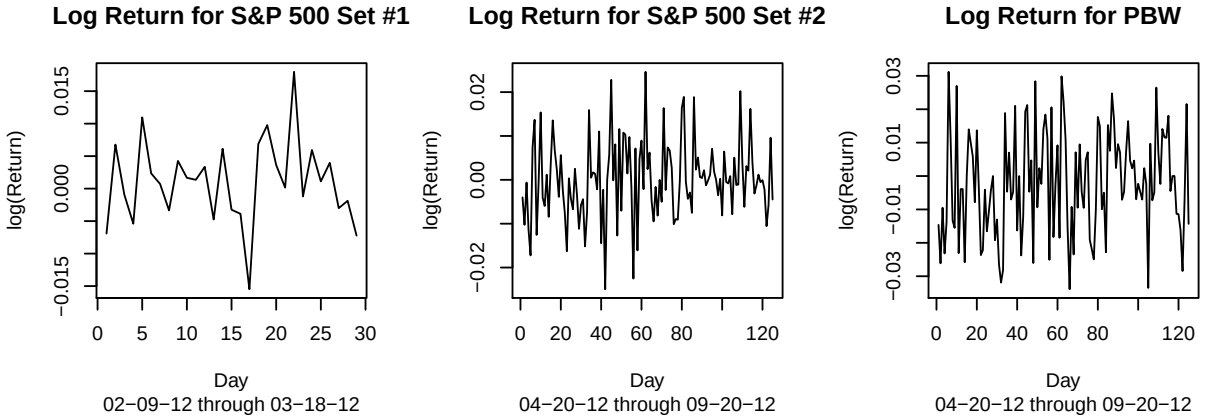

Figure 4.5: Log Returns After Transform

Figure 4.6: Log Returns After Transform

daily score set for the S&P500 (Figure 4.7, center). It can be seen that Apple Set #1 (Figure 4.8) is skewed toward positive sentiment. In the time period during which collection took place, Apple saw a significant rise in price and became the largest company in the world by market capitalization. The proposed framework is promising because this rise and positive coverage is evident in the histogram. The second Apple set was during the subsequent retraction after the stock experienced continual record highs.

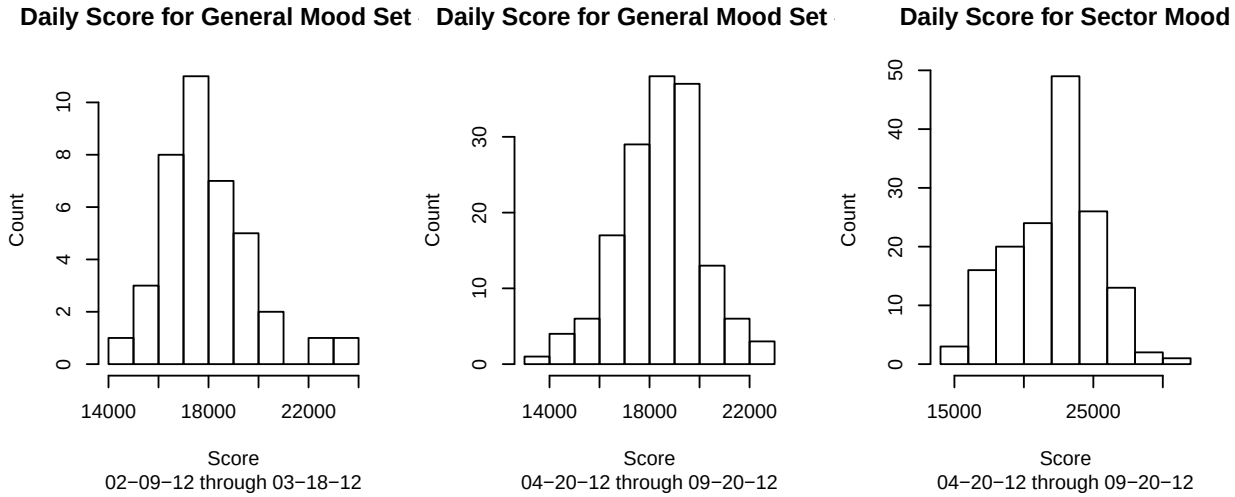

Figure 4.7: Histograms of Daily Score

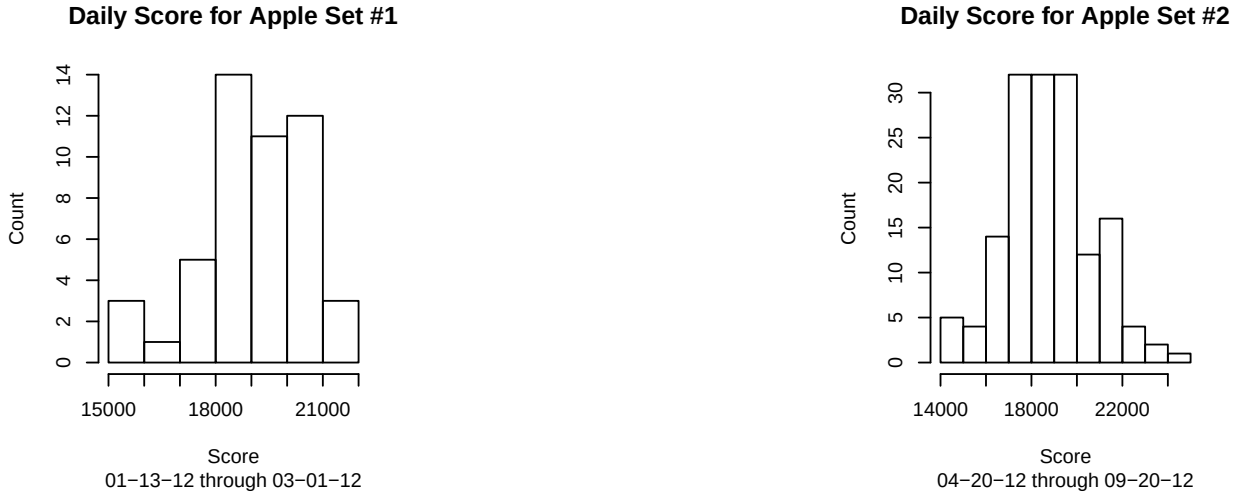

Figure 4.8: Histograms of Daily Score

4.1 Linear Regression

To create a model for prediction, the analysis began with a multiple regression. This regression was used in order to see which lags could be a significant indicator of price. From past research, lags past lag-5 are not relevant to prediction [4]. Lagging the data also decreases

the size of the data set. Due to limited data and constraining effects of the lagging process, data were lagged a maximum of 5 days. Thus, the data set contained the log returns of the stock, along with lag-1, lag-2, lag-3, lag-4, and lag-5 returns with lag-0, lag-1, lag-2, lag-3, lag-4, and lag-5 scores.

There is also evidence that the volume of trading can be influential in determining the movement of a stock. To make a more complete model, the log of the volume up to lag-5 was also included. Thus, the data set contained the log volume of the stock, lag-0, lag-1, lag-2, lag-3, lag-4, and lag-5 volumes.

The full model for the sets tested for all areas is below:

$$\begin{aligned}\hat{P} = \beta_0 + \beta_1 R_{t-1} + \beta_2 R_{t-2} + \beta_3 R_{t-3} + \beta_4 R_{t-4} + \beta_5 R_{t-5} + \\ \beta_6 S_t + \beta_7 S_{t-1} + \beta_8 S_{t-2} + \beta_9 S_{t-3} + \beta_{10} S_{t-4} + \beta_{11} S_{t-5} + \\ \beta_{12} V_t + \beta_{13} V_{t-1} + \beta_{14} V_{t-2} + \beta_{15} V_{t-3} + \beta_{16} V_{t-4} + \beta_{17} V_{t-5}\end{aligned} \tag{4.2}$$

Where $\hat{P}$ is the predicted log-return, R is the appropriately lagged return for time t, S is the appropriately lagged score for time t, V is the appropriately lagged volume for time t, and β is the corresponding coefficient.

Stepwise selection based on Akaike Information Criterion (AIC) was used to find the model which presented the best tradeoff between complexity and reliability. Stepwise selection through AIC starts with a null model and adds and removes multiple combinations of variables using AIC for the entire tested model as a comparative method. The model with the lowest AIC is chosen and that model is used.

The models for all sets show a fairly significant linear component. Market and sector-wide

Table 4.1: AIC Model Selection Results

Area Explored	Selected Model	Coefficient Values	AIC	Model p-value
Market Wide 1st	$\beta_0 + \beta_1 S_{t-3}$	$\beta_0 = 0.0263$, $\beta_1 = -7.77 \times 10^{-7}$	-172.54	0.08525
Market Wide 2st	$\beta_0 + \beta_1 R_{t-4}$ $+\beta_2 S_{t-5}$	$\beta_0 = 0.0148$, $\beta_1 = -0.01945$ $\beta_2 = -1.41 \times 10^{-6}$	-795.54	0.0277
Sector Wide Set	$\beta_0 + \beta_1 S_{t-5}$	$\beta_0 = 0.0177$, $\beta_1 = -7.27 \times 10^{-7}$	-656.45	0.13900
Apple 1st Set	$\beta_0 + \beta_1 S_{t-1}$	$\beta_0 = 0.0455$, $\beta_1 = -2.75 \times 10^{-6}$	-232.47	0.07941
Apple 2nd Set	$\beta_0 + \beta_1 S_{t-4}$	$\beta_0 = 0.0318$, $\beta_1 = -1.68 \times 10^{-7}$	-606.65	0.0659

sets show near significance, but do not break the desired threshold of $\alpha = .10$. As sentiment score for the respective tested area increases, the log-return for the respective area does as well. Sector and Market sets show a 3-5 day lag similar to that of previous studies. This shows a promising correlation and the lag makes intuitive sense. Market participants do not absorb data instantaneously and there is a delay between emotion and action.

The Apple #1 and Apple #2 both produce significant linear components. In both cases, the relationship between score and returns for Apple happen between a shorter time frame than the corresponding Market or Sector sets. A possible explanation is that Apple's user base is more tech savvy than the general populous, therefore investors would be able to react to news more quickly than the whole population of investors seen in the S&P500.

Testing the overall market sentiment and its influence on the sector and company wide data sets is also of interest. The predicted price should not only be influenced by the score being tested, but also by the score for the market as a whole. Thus, the testing for company and sector influence also includes market-wide scores and volume. The smaller market and company data sets were omitted due to lack over overlapping data points.

The full model for the company and sector sets tested for all areas is below:

$$\begin{aligned}\hat{P} = \beta_0 + \beta_1 R_{t-1} + \beta_2 R_{t-2} + \beta_3 R_{t-3} + \beta_4 R_{t-4} + \beta_5 R_{t-5} + \\ \beta_6 S_t + \beta_7 S_{t-1} + \beta_8 S_{t-2} + \beta_9 S_{t-3} + \beta_{10} S_{t-4} + \beta_{11} S_{t-5} + \\ \beta_{12} V_t + \beta_{13} V_{t-1} + \beta_{14} V_{t-2} + \beta_{15} V_{t-3} + \beta_{16} V_{t-4} + \beta_{17} V_{t-5} + \\ \beta_{18} M_t + \beta_{19} M_{t-1} + \beta_{20} M_{t-2} + \beta_{21} M_{t-3} + \beta_{22} M_{t-4} + \beta_{23} M_{t-5} + \\ \beta_{24} Q_t + \beta_{25} Q_{t-1} + \beta_{26} Q_{t-2} + \beta_{27} Q_{t-3} + \beta_{28} Q_{t-4} + \beta_{29} Q_{t-5}\end{aligned} \tag{4.3}$$

Where $\hat{P}$ is the predicted price, R is the appropriately lagged return for time t, S is the appropriately score for time t, V is the appropriately lagged volume for time t, M is the appropriately lagged market log return for time t, Q is the appropriately lagged market volume for time t, and β is the corresponding coefficient.

The AIC stepwise selection method selects the same model as before, yielding significant

Table 4.2: AIC Model Selection Results

Area Explored	Selected Model	Coefficient Values	AIC	Model p-value
Sector Wide Set	$\beta_0 + \beta_1 S_{t-5} + \beta_2 M_{t-5}$	$\beta_0 = 0.0068$, $\beta_1 = 8.42 \times 10^{-6}$, $\beta_2 = -1.47 \times 10^{-7}$	-657.34	0.08227
Apple 2nd Set	$\beta_0 + \beta_1 S_{t-4}$	$\beta_0 = 0.0318$, $\beta_1 = -1.68 \times 10^{-7}$	-606.65	0.0659

results. It is interesting that the lowest AIC was the same for both Apple sets. Selection of the same lag over different time periods for Apple may still indicate a meaningful relationship. The Sector Wide Set was just under the $\alpha = .10$ threshold for significance, and does include a lag similar to that of the market as tested above. Market sentiment combined with sector sentiment could influence sector investment.

4.2 Time Series Component

There lies a potential for incorporating a time series component into the above model. If there is a temporal component to the data, then the linear regression model would not be

appropriate. Under these circumstances, there will be a pattern in traditional regression diagnostics.

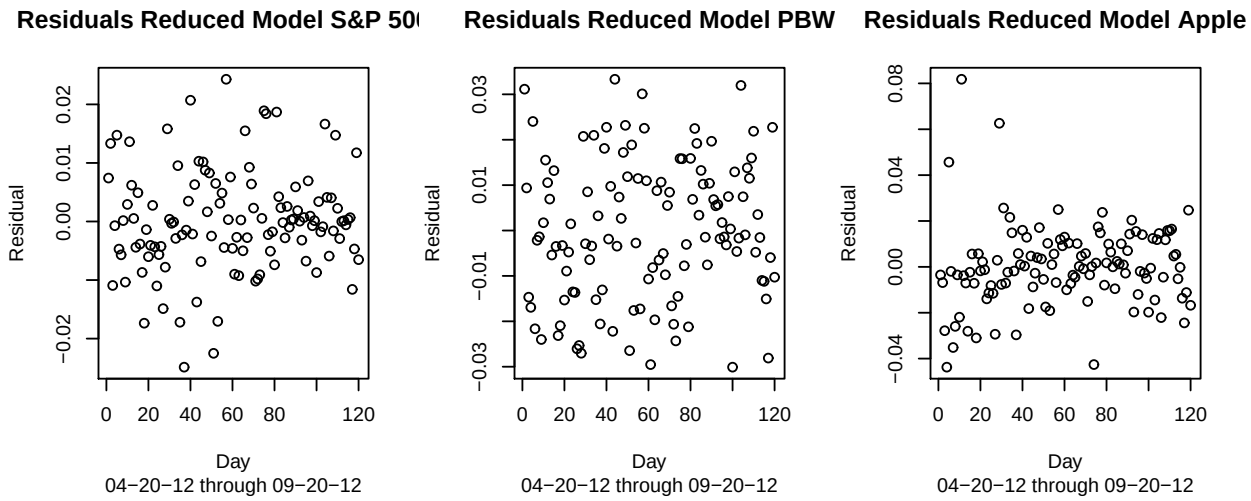

Figure 4.9: Residual Plots for Reduced Models

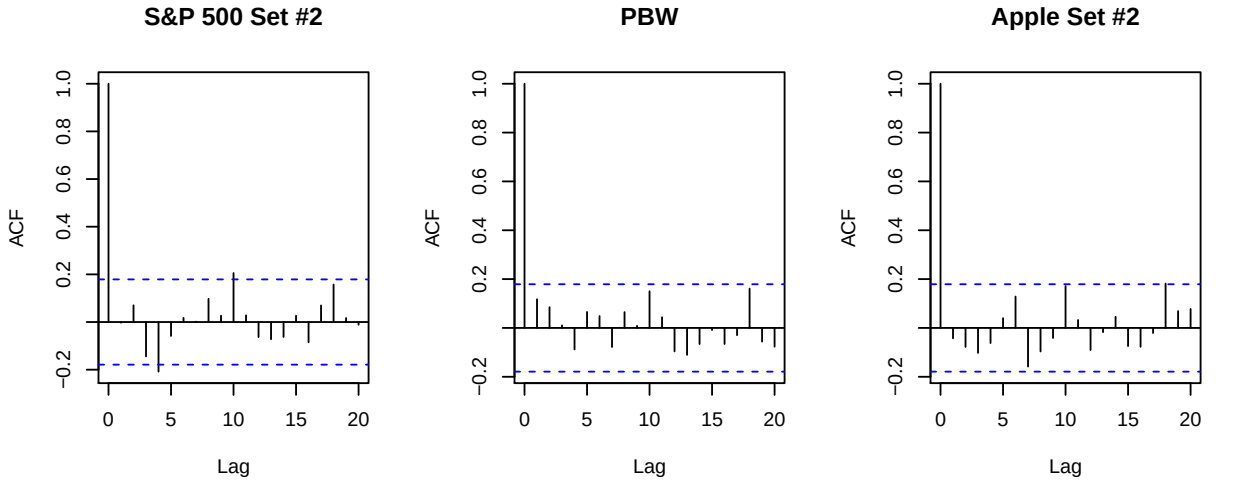

Figure 4.10: ACF of Residuals for Reduced Models

The residual plots, in Figure 4.9 produce patterns heavily clustered around 0, and the ACF

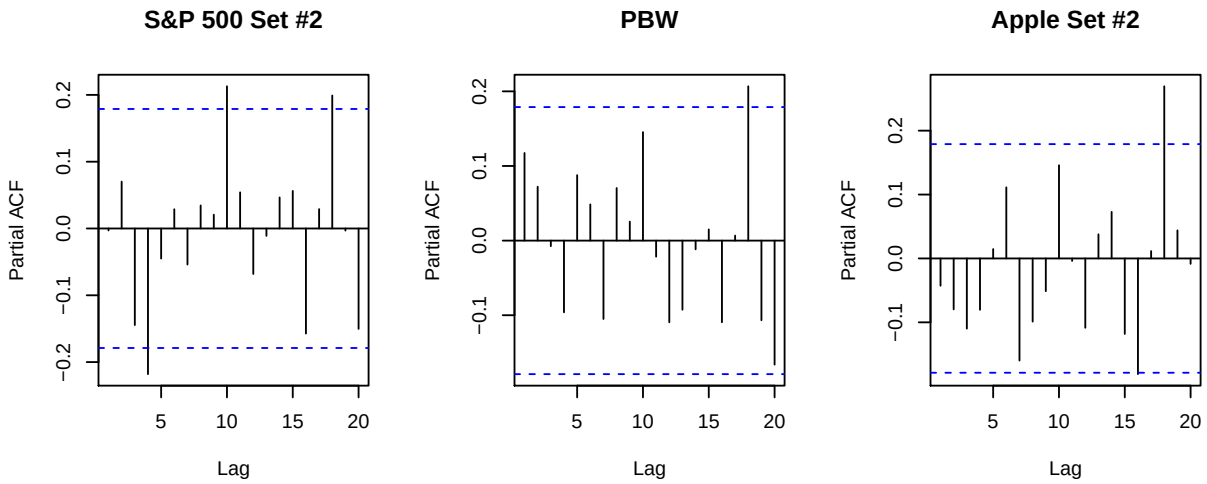

Figure 4.11: PACF of Residuals for Reduced Models

(Figure 4.10) and PACF (Figure 4.11) plots of the Market Set exhibit wavelike patterns and has a significant coefficient at lag-4, the same as the linear model above. The models are expanded to model the linear component as well as a relationship of the errors. If a temporal correlation is present in the data, then using a linear regression will only account for the linear component, leaving serially correlated errors. The errors can be further modeled in addition to the linear regression model with the following formula:

$$\hat{y} = \beta_0 + \beta_1 R_t + ... + \beta_m R_{t-m} + \epsilon_{t-m} + \rho \tag{4.4}$$

Where the linear component is present with lagged scores and coefficients, ϵ is the lagged error from the specified previous day, and ρ is the shock parameter. The shock parameter is what is typically thought of as error in a model. ρ should be normally distributed with mean of 0 and variance of 1. Once modeled in this way, the regression component accounts for the linear component and the error term accounts for the time series component. Since the Market Set was the only set with both significant linear and temporal components, it is the

only set tested henceforth. Adding in a time series component doesn't helpfully contribute

Table 4.3: AIC Model Selection Results

Area Explored	Selected Model	AIC	Model p-value
Market Wide 2nd Set	$\beta_0 + \beta_1 R_{t-4} + \beta_2 S_{t-5} + \epsilon_{t-4} + \rho$	-790.66	0.1187

to the model. An increase in AIC and negligible change in p-value does not add enough information to warrant the complexity of its inclusion.

4.3 Granger Causality Test

The linear regression model shows correlation, but not direction of that correlation, although it is implied via the lags due to the time component. For model validation of the above models, a Granger causality test was used. This test, indicating potential direction of causality on one time series from another, shows the linear regression models above were correct in variable selection. This test indicates when one time series is useful in predicting another. It shows, with these sets, that there is a significant effect of predicting log returns from lagged score. Table 4.4 shows the results of a Granger test run in each direction on all possible combinations of log returns and each lagged score. The lags are similar to those of the linear regression, and show that there is evidence of one time series influencing another.

Table 4.4: Granger Causality Results

Area Explored	Significant Lag	F-Statistic	Granger Causality p-value
Market Wide 1st Set	Lag-3	3.2140	0.08814
Market Wide 2st Set	Lag-5	2.5058	0.11614
Sector Wide Set	Lag-5	2.1043	0.14957
Apple Set #1	Lag-1	3.96887	0.05320
Apple Set #2	Lag-4	3.40114	0.06770

4.4 Logistic Regression

Predicting direction and degree of log returns may be too ambitious, so a method to predict only direction was tested. This test was based on the log returns where positive returns, indicative of the price going up, were given an indicator of 1. The rest, or the negative log returns, were given an indicator of 0. The same continuous inputs were used, and the model was the following:

$$\begin{aligned}\log\left[\frac{\theta(\hat{p})}{1-\theta(\hat{p})}\right] = \beta_0 + \beta_1 R_{t-1} + \beta_2 R_{t-2} + \beta_3 R_{t-3} + \beta_4 R_{t-4} + \beta_5 R_{t-5} + \\ \beta_6 S_t + \beta_7 S_{t-1} + \beta_8 S_{t-2} + \beta_9 S_{t-3} + \beta_{10} S_{t-4} + \beta_{11} S_{t-5} + \\ \beta_{12} V_t + \beta_{13} V_{t-1} + \beta_{14} V_{t-2} + \beta_{15} V_{t-3} + \beta_{16} V_{t-4} + \beta_{17} V_{t-5}\end{aligned} \tag{4.5}$$

Where $\theta(\hat{p})$ is the probability if success, R is the appropriately lagged return for time t, S is the appropriately lagged score for time t, V is the appropriately lagged volume for time t, and β is the corresponding coefficient.

The models produced by stepwise AIC selection indicate a similar lag to the linear models.

Table 4.5: AIC Model Selection Results

Area Explored	Selected Model	Coefficient Values	AIC	Model p-value
Market Wide 2st Set	$\beta_0 + \beta_1 S_{t-4} + \beta_2 S_{t-5}$	$\beta_0 = 11.6451$, $\beta_1 = -0.00036$ $\beta_2 = -0.00024$	156.82	0.00045
Sector Wide Set	$\beta_0 + \beta_1 S_{t-3}$	$\beta_0 = 1.037$, $\beta_1 = -6.242 \times 10^{-5}$	166.07	0.337
Apple 2nd Set	$\beta_0 + \beta_1 S_{t-2} + \beta_2 M_{t-5}$	$\beta_0 = 9.4356$, $\beta_1 = -0.0002$, $\beta_2 = -0.0002$	161.90	0.0061

The Market set and Apple sets are significant. Notably, the Apple direction of price includes an Apple score component as well as a Market component. This is indicative of Apple's direction being influenced not only its sentiment, but the Market as a whole. This also is consistent with financial theory and both Apple sets from the previous section wherein the

lag time between score and Apple returns is less than the Market set.

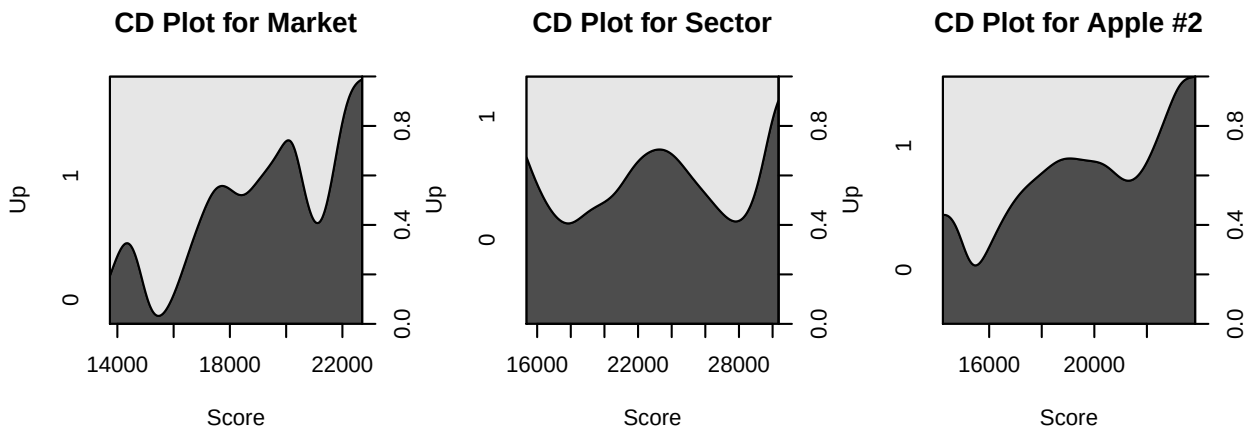

Figure 4.12: Cumulative Density Plots for Logistic Reduced Models

The cumulative density plots, shown in Figure 4.12, using the best single parameter model selected by AIC, show a binary relationship between score and direction. These plots show the expected outcome y, for a given score x. In this case, 1 represents up and 0 represents down. The Market set has the best relationship, while the Sector set has no relationship.

To test model reliability, Receiver Operating Characteristic (ROC) Plots, in Figure 4.13, were generated to further examine the models. Logistic regression predictions output probabilities instead of binary results. To transform the logistic regression predictions into binary, a probability cutoff for the predicted model must be selected. ROC plots map out, on a continuous line, the relationship between true positives and false positives.

The ideal model on an ROC plot would be a line from (0,0) to (1,0) on the coordinate plane. This would mean that the model is always accurate in its binary prediction. A line going from (0,0) to (0,1) would indicate a model that is no better than even odds. Because the Market model is the closest to (1,0) on the coordinate plane, it has the best true positive

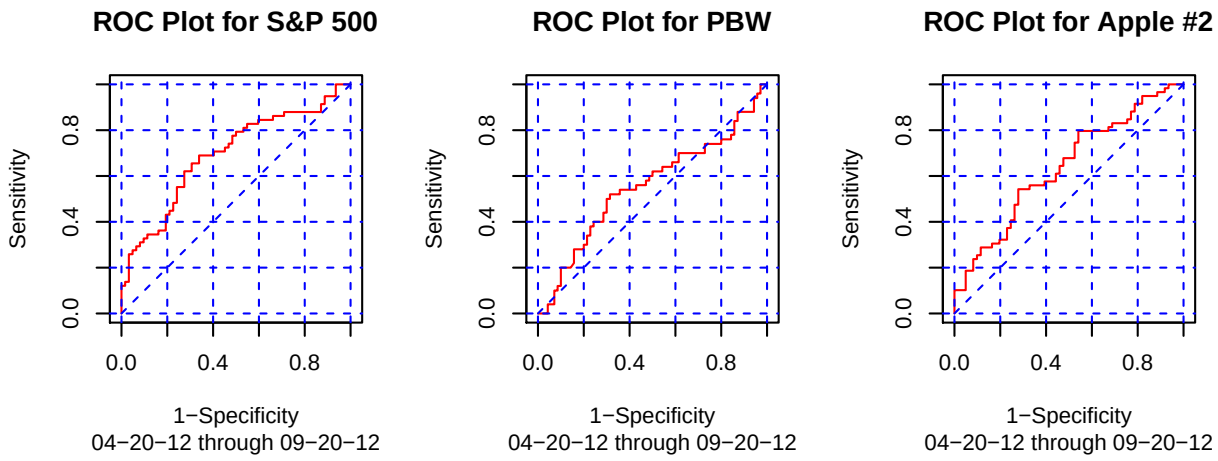

Figure 4.13: Receiver Operating Characteristic Plots for Logistic Reduced Models

rate out of the three tested areas of interest.

Choosing a probability threshold from the ROC plot is somewhat subjective and is dependent on the purpose of the model with regards to how harmful false positives are. In the case of predicting log returns, a false positive means that a buy or sell signal was indicated, when the market truly goes down or up, respectively. It is quite important to minimize losses with a market prediction model, thus a lower false positive threshold is desired. Fortunately, all three models have an almost vertical point, meaning that there is a significant increase in true positives for a negligible increase in false positives. This point appears to be around $P[FalsePositive] = .20$ for all models. A desired rate of false positives could also be selected based on risk preference.

To test the success of the model, the predictions for every point were generated. The predictions where then given a value of 0 or 1 based on $P[FalsePositive] = .20$ above. This probability was chosen to minimize false positives. The following table contains the accuracy

of each model.

Table 4.6: AIC Model Selection Results

Area Explored	F-Positive	T-Positive	F-Negative	T-Negative	Total Correct (%)
Market Wide 2st Set	17	36	22	45	67.5
Sector Wide Set	0	0	50	70	58.3
Apple 2nd Set	22	33	26	39	60

Chapter 5

Conclusions

This work has been rather inconclusive. The negative coefficients, but positively correlated ROC and cumulative density plots offer conflicting views of the underlying truth. What this work has done is shown that a more refined scoring methodology is necessary the prediction of stock market returns. Through Twitter, gauging user sentiment is a very real possibility, as shown by other work. The public perception of a company can ultimately influence an investors decision to buy that particular stock. While promising, more work needs to be done to test the scoring methodology further. The models selected through AIC were also too simple to be taken as the best possible model. It is known that volume and previous returns have at least some influence on future returns. A more complex statistical approach coupled with a more inclusive model may aid in finding a sentiment component to stock market prediction.

Because the idea of using Twitter to model events is in its infancy, an untold number of approaches remain untested. First, semantic analysis of large streaming data is hardly evolved. The approach proposed in this thesis is heavily reliant on word counting, which leaves it not very robust against manipulation. Unscrupulous users could spam positive or negative keywords along with a company name and render this approach meaningless. More

work in developing semantic analysis of this type is necessary. Creation of a stronger lexicon would greatly improve the proposed approach. This lexicon relies on about 1450 words, most of which are positive. The lexicon also does not include slang, acronyms, or swear words. Speech patterns on Twitter include all of these features, so including them should improve the reliability of models, as well as a better indication of sentiment.

More importantly, further work into a stopwords lexicon would vastly improve processing speed. Quite a few words are allowed by this methodology which are routinely used and offer no sentence influence. Increasing the proportion of filtered words would decrease total words to be analyzed and produce noticeable speed increases. A better methodology for sampling needs to be developed. Twitter offers access to 1.0%, 10%, and 100% systematic samples, while this methodology systematically sampled every minute. No work has been done on assessing other potential sampling methods. Improvements in this area may give a more robust snapshot of the population or significantly decrease processing time.

References

[1] E. Naone, *What Twitter Learns from All Those Tweets,* 28 09 2010. [Online]. Available: http://www.technologyreview.com/blog/editors/25809/. [Accessed 03 06 2012].

[2] E. F. Fama, *The Behavior of Stock-Market Prices,* The Journal of Business, vol. 38, no. 1, pp. 34-105, 1965.

[3] B. Qian and R. Khaled, *Stock Market Prediction with Multiple Classifiers,* Applied Intelligence, vol. 26, no. 1, pp. 13-23, 2007.

[4] J. Bollen, H. Mao and X.-J. Zeng, *Twitter Mood Predicts the Stock Market,* Journal of Computational Science, pp. 1-8, 2011.

[5] K. Kim and J. Nofsinger, *The Behavior of Japanese Individual Investors During Bull and Bear Markets,* The Journal of Behavioral Finance, vol. 8, no. 3, pp. 1-16, 2007.

[6] W. F. M. D. Bondt and R. Thaler, *Does the Stock Market Overreact?,* The Journal of Finance, vol. 40, no. 3, pp. 793-805, 1986.

[7] M. Cheong and V. Lee, *A Microblogging-Based Approach to Terrorism Informatics: Exploration and Chronicling Civilian Sentiment and Response to Terrorism Events via Twitter,* Journal of Information Systems Frontiers, vol. 13, no. 1, pp. 45-59, 2011.

[8] C. Oh and O. R. L. Sheng, *Investigating Predictive Power of Stock Micro Blog Sentiment in Forecasting Future Stock Price Directional Movement,* in Thirty Second International Conference on Information Systems, Shanghai, 2011.

[9] S. Asur and B. A. Huberman, *Predicting the Future With Social Media,* 2010. [Online]. Available: http://www.hpl.hp.com/research/scl/papers/socialmedia/socialmedia.pdf. [Accessed 2012].

[10] S. A. Golder and M. W. Macy, *Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures,* Science Magazine, pp. 1878-1881, 30 September 2011.

[11] M. Guy, P. Earle, C. Ostrum, K. Gruchalla and S. Horvath, *Integration and Dissemination of Citizen Reported and Seismically Derived Earthquake Information via Social Network Technologies,* U.S. Geological Survey National Earthquake Information Center, Golden, 2010.

[12] S. Abidin, N. Omar, M. Radzi and M. Haron, *Quantifying Text-based Public's Emotion and Discussion Issues in Online Forum,* The Society of Digital Information and Wireless Communications, 2011.

[13] W. Antweiler and M. Z. Frank, *Is All That Talk Just Noise?,* The Journal of Finance, vol. 59, no. 3, pp. 1259-1294, 2004.

[14] G. Miller, *Social Scientists Wade Into the Tweet Stream,* Science Magazine, pp. 1814-1815, 30 September 2011.

[15] E. F. Fama, *Market Efficiency, Long-Term Returns, and Behavioral,* Journal of Financial Economics, vol. 49, pp. 283-306, 1998.

[16] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski and L. Brilliant, *Detecting Influenza Epidemics Using Search Engine Query Data,* Nature, vol. 457, pp. 1012-1014, 19 Februrary 2009.

[17] C. Basu, *Investment Performance of Cormnon Stocks in Relation to Their Price-Earnings Ratios: A Test of the Efficient Market Hypothesis.* The Journal of Finance Volume 32, No. 3, June 1977, pp. 663-682.

[18] Moorad Choudhry, Stuart Turner, Gino Landuyt and Khurram Butt, *Modern Portfolio Theory and the Myth of Diversification* World Commerce Review Vol. 3, Issue 1 2009

[19] Fama and French, *The Cross-Section of Expected Stock Returns* The Journal of Finance Volume 67, 1992, pp. 427-465.

[20] http://www.momswhothink.com/reading/list-of-adjectives.html

[21] http://www.momswhothink.com/reading/list-of-adverbs.html

[22] Granger, Clive W.J., Huang, Bwo-Nung, Yang, Chin W. *A Bivariate Causality between Stock Prices and Exchange Rates: Evidence from Recent Asia Flu,* The Quarterly Review of Economics and Finance, vol. 40, no. 3, 2000, pp. 337-354.