STOCK PREDICTION ANALYZING INVESTOR SENTIMENTS

A Dissertation
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Arijit Chatterjee

In Partial Fulfillment of the Requirements
for the Degree of
DOCTOR OF PHILOSOPHY

Major Department:
Computer Science

May 2017

Fargo, North Dakota

North Dakota State University
Graduate School

Title

STOCK PREDICTION ANALYZING INVESTOR SENTIMENTS

By

Arijit Chatterjee

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

DOCTOR OF PHILOSOPHY

SUPERVISORY COMMITTEE:

Dr. Kendall Nygard
Chair

Dr. Gursimran Walia

Dr. Saeed Salem

Dr. Amitava Chatterjee

Approved:

March 10, 2018
Date

Dr. Kendall Nygard
Department Chair

ABSTRACT

We are going through a phase of data evolution where a major portion of the data from our daily lives is now been stored on social media platforms. In recent years, social media has become ubiquitous and important for social networking and content sharing. Sentiment analysis and opinion mining is the field of study that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written language.

In the financial sector, sentiments are also of paramount importance, and this dissertation mainly focuses on the effect of sentiments from investors [3] on the behavior of stocks. The dissertation work leverages social data from Twitter and seeks the sentiment of certain investors. Once the sentiment of the tweets is calculated using an advanced sentiment analyzer, it is used as an additional attribute to the other fundamental properties of the stock. This dissertation demonstrates how incorporating the sentiments improves forecasting accuracy of predicting stock valuation. In addition, various experimental analysis on regression based statistical models are considered which show statistical measures to consider for effectively predicting the closing price of the stock. The Efficient Market Hypothesis (EMH) states that stock market prices are largely driven by additional information and follow a random walk pattern [7, 8, 37, 39, 41]. Though this hypothesis is widely accepted by the research community as a central paradigm governing the markets in general, several people have attempted to extract patterns in the way stock markets behave and respond to external stimuli. We test a hypothesis based on the premise of behavioral economics, that the emotions and moods of individuals basically the sentiments affect their decision-making process, thus, leading to a direct correlation between “public sentiment” and “market sentiment” [42, 43, 44, 45]. We first select key investors from Twitter [27, 28] whose sentiments matter and do sentiment analysis on the tweets pertaining to stock related information.

Once we retrieve the sentiment for every stock, we combine this information with the other fundamental information about stocks and build different regression-based prediction models to predict their closing price.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude and appreciation for my adviser and research supervisor, Dr. Kendall Nygard, for his strong support and continuous guidance which encouraged me to strive for the highest level of achievement by performing this work to my fullest, successfully resulting in several scholastic and research accomplishments. Additionally, special acknowledgements go to all other supervisory committee members, Dr. Gursimran Walia, Dr. Saeed Salem and Dr. Amitava Chatterjee, for exclusively dedicating their valuable time and attention for my own sake. I express my sincere gratitude to all other faculty members from the Computer Science Department to whom I am deeply grateful for teaching me through their valuable courses to fulfill my relentless dream of earning the doctorate degree and to our department secretary Carole Huber for always supporting me.

My sincere prayers to Babaji and Gopal without whose grace I could never achieve any of my accomplishments. I am grateful to my parents, Suparna and Dr. Mihir Baran Chatterjee, who did not spare any effort in raising me the way I am now, to my late grandparents Mr. Indu Shekhar Chatterjee, Mrs. Gauri Rani Chatterjee, Mr. Binoy Kumar Bhattacharjee and Mrs. Anita Bhattacharjee who always motivated me to realize my childhood dreams. I am thankful to my beloved wife Sudeshna who has always been supportive and encouraging, to my in-laws Dr. Ashoke and Dr. Maya Ray for their advice and to all my close and dear family members and friends who have been a part of this remarkable journey.

Last but not the least, I would like to thank Microsoft Research and Development for sponsoring my PhD and providing me with all possible resources to help me achieve this degree.

DEDICATION

To the memory of my beloved late research adviser, Dr. William Perrizo, whom I miss every day and without whom I wouldn't even dream of pursuing a doctorate degree. His passion for research and love for academics will always be carried forward with me. I owe my doctorate degree entirely to him.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

Sentiment is a view of or attitude toward a situation or event. It is an opinion, feeling on a subject and depicts the emotion involved. Sentiment analysis systems are being applied in almost every business and social domain because opinions are central to almost all human activities and are key influencers of our behaviors. Our beliefs and perceptions of reality, and the choices we make, are largely conditioned on how others see and evaluate the world. For this reason, when we need to decide, we often seek out the opinions of others. This is true for individuals and for organizations.

Sentiments are a key element now to understand the relevant information and is an upcoming trend in the field of processing information. Sentiment Analysis is performed on tweets, posts, and news articles to understand trends and to make decisions based on them. Launched in the year 2006, Twitter has become one of the most commonly used social networking sites. But with so much data collected in Twitter [85] or other Social Media platforms, it is virtually impossible to keep track of just the information in which we are interested. For mining these data efficiently for decision making or decision support, it requires fast processing of those data with an acceptable degree of accuracy [1, 2].

Predicting the stock market is not a simple task. Mainly because of the close to random-walk behavior of a stock time series [19]. Different techniques are being used in the trading community for prediction tasks. There are also several attributes which are considered to improve the prediction of stock price. In recent years, the concept of sentiment analysis [9,10] has emerged as one of them. In this dissertation, we investigate and empirically evaluate how sentiment analysis from social media content of key investors can be used as an additional parameter [29, 30] to the fundamental properties of the stock like open price, close price, high price, low price, volume of

stocks traded for predicting real-world outcomes such as the close price of a stock on the following day. We also show how Azure sentiment analyzer outperforms other commercial sentiment analysis tools and how we calculate the daily average sentiment of 3000 chosen investors pulling the tweet data from Twitter. We use the daily calculated "AvgSentiment" score as an additional property of every stock and do statistical studies to show the effect of this score on the close price of the stock. We perform evaluations of the time periods for which sentiments affect the close price of the stock and use our analysis to build various regression-based prediction models. We show that with an acceptable degree of accuracy and significant coefficient of determination, we can predict the close price of the stock on the following day using the historical stock information.

The brief outline of this dissertation is as follows. A literature review of the related work and details of the published papers which comprise this dissertation is described in Chapter 2. An overview of Twitter as a social platform and the effect of investor sentiments and building the Azure Sentiment Analyzer and comparing with the other NLP tools is discussed in Chapter 3. Building different regression-based prediction models using sentiment score as an attribute is discussed in Chapter 4, and we finally summarize and conclude this dissertation in Chapter 5 by presenting some of the assumptions we have considered and wrapping up the contributions of our work.

CHAPTER 2. LITERATURE REVIEW

Sentiment describes a group of people's opinions, emotions or views. Investor sentiment is an approach to measure market sentiment. Dreman et al. [49] in their research mention about how investor sentiment surveys have long displayed interesting investor attitudes over the years. For example, despite large April 2000 market losses, investor expectations of future returns did not significantly fall. Thorp [50] mentions that extremely bullish levels of sentiment often come after strong market run-ups when investors are fully invested in the market. While Malkiel and Fama's (1970) Efficient Market Hypothesis (EMH) indicates that securities prices fully reflect all publicly available information, in 2003 Shiller et al. [51] finds that, "The efficient markets model, for the aggregate stock market, has still never been supported by any study effectively linking stock market fluctuations with subsequent fundamentals." In 2001, Hall [52] adds that many high-tech companies with negative earnings maintained high stock prices for long periods of time. In 2003, Malkiel [53, 54] and other economists challenged the EMH, to explain diverging market sentiment, including how psychological and behavioral elements impact stock prices. Some advocates even recommended using investor sentiment as a contrarian indicator for the overall market in certain specific situations.

2.1. Investor Sentiment Measurements and Effect on the Stock Market

Barberies et al. [55] in their research study show investor sentiment associated unreliably with stock prices. They showed how interestingly, investor sentiment under-reacted to more factual information such as earnings announcements, share repurchases, dividend initiations, and overreacted to a prolonged record of extreme (good or bad) performances. However, in 2006 Baker and Wurgler [7, 8] conclude that "…waves of sentiment have clearly discernible, important, and regular effects on individual firms and on the stock market as a whole." In 2004, Thorp [50] also

comments on the lagging feature of sentiment as well as the potential irrational emotions that drive prices. He notes that "week to week changes in member sentiment do not reveal meaningful relationships between sentiment and market performance," but he does discover that excessive investor sentiment in either a bullish or bearish direction would signal a significant opposing response over the following 52 weeks. Several studies find some measures of investment sentiment predicting stock returns. In 2006, Lemmon and Portniaguina [56] find that investor sentiment forecasts the returns of small stocks. In 2015, Zheng [57] documents a negative predictive relation between sentiment and metal futures' returns. In 2015, Kaplanski et al. [58] affirm that more positive sentiment associates with higher return expectations and higher intentions to buy stocks. In 2014, Ling et al. [59] find a positive association between investor sentiment and subsequent private market returns. In 2015, Babu and Kumar [60] document that negative sentiment has a greater bearing on the NSE index return than positive sentiments.

In 2012, Mittal and Goel [14] used sentiment analysis and machine learning principles to find the correlation between "public sentiment" and" market sentiment" to predict public mood and use the predicted mood and previous days' DJIA values to predict the stock market movements. In 2010, Bollen, Mao and Zeng [11] investigated whether measurements of collective mood states derived from large-scale Twitter feeds are correlated to the value of the Dow Jones Industrial Average (DJIA) over time. They have analyzed the text content of daily Twitter feeds by two mood tracking tools, namely OpinionFinder that measures positive vs. negative mood and Google-Profile of Mood States (GPOMS) that measures mood in terms of 6 dimensions (Calm, Alert, Sure, Vital, Kind, and Happy). In 2015, Chatterjee and Perrizo [61] described how selecting tweets from 3000 key investors from Twitter can be a good indicator of stocks which are most frequently been discussed. They also make careful considerations that an investor can be biased

on a particular stock and how this biasness can affect the volatility of the stock in the market. The above research literature finds a substantial body of research of significant associations between investor sentiment and market return related measures.

2.2. Investor Biasness

In 2004, Shefrin [62, 63] denotes three behavioral finance themes: heuristic-driven bias, frame dependence, and inefficient markets. Heuristics call problem solving techniques "rules of thumb and not strict logic." These often simple and efficient mental process rules tend to help people with decision making; a limited focus could lead to errors from cognitive bias. Framing is the process to see the world with all of our mental and emotional filters from our past experiences. Inefficient markets denote the price or rate of return that appears to contradict the EMH. Overconfidence becomes important when analyzing market inefficiencies. Psychology studies generally agree that human beings tend to overestimate their abilities. Regarding financial decisions, in 1999, Barber and Odean [64] present that investors' overconfidence creates bias in their abilities and in regretting poor decisions. The Psychological - Overconfidence Theory indicates that investors tend to overweight private information while ignoring public information. In 2006, Chung and Lee [65] examined the impact of public and private information shock on trading volume and equal-weighted stock prices, finding that value weighted stock prices strongly overreact to private information shock and under react to public 4 information shock. Also, private information has an earlier impact than public information on stock prices and equal-weighted stock prices under-react to a public information shock for a longer period. Thus, overconfident investors trade more aggressively in subsequent periods when making market gains. Their findings are consistent with the expectation of the overconfidence hypothesis. In 2016, Chatterjee [66] evaluate the sentiment of the key investors they have identified to study their effect on certain ticker

symbols. Their work references the *top down* approach from 2006 by Baker and Wurgler [7, 8] which focuses on the measurement of reduced form, aggregate sentiment and traces its effects to market returns and individual stocks. The approach is based on two broad undisputable assumptions of behavioral finance—sentiment and the limits to arbitrage—to explain which stocks are likely to be most affected by sentiment, rather than simply pointing out that the level of stock prices in the aggregate depends on sentiment. In particular, stocks of low capitalization, younger, unprofitable, high volatility, non-dividend paying, growth companies, or stocks of firms in financial distress, are likely to be disproportionately sensitive to broad waves of investor sentiment. Stocks that are difficult to arbitrage or to value are most affected by sentiment. Chatterjee and Perrizo [66] also show when sentiment is low, the average future returns of speculative stocks exceed those of bond-like stocks. When sentiment is high, the average future returns of speculative stocks are on average lower than the returns of bond-like stocks. In their analysis, they focus on investors being biased on ticker symbols and would remove any bias in their tweets based on their tweet counts. The more is the heterogeneity of the population of investors [4] tweeting on ticker symbol T_s, signifies T_s is been discussed by a wider population range.

2.3. Sentiment Analysis and Approaches

The recent data explosion has spawned an incredible increase in innovation. Sentiment analysis is a newer field that has only recently traversed from the academic realm to corporate use. Much of the current published research on the subject was developed by research facilities strongly associated with companies such as IBM, Microsoft, Google. “The sentiment detection of texts has witnessed a booming interest in recent years” (in 2009, Tang et al. [67]) with “the emergence of new social media such as tweets, blogs, message boards, news, and web content” dramatically changing the ecosystems of corporations (in 2010, Cai et al. [68]). The academic

contributors to the subject have combined many specific areas of linguistics, computer science, artificial intelligence, and psychology. More specifically as mentioned by Tang et al. [67] in 2009, it is “a discipline at the crossroads of NLP [natural language processing] and IR [information retrieval], and as such it shares a number of characteristics with other tasks such as information extraction and text-mining". Machine learning techniques, basic statistical analysis, and linguistic semantic representation are also well represented in the designs of the field. As with many new fields, sentiment analysis is a combination of a few novel concepts reapplied to a wide range of specific aspects of other older fields. In 2010, Cai et al. [68] describes this importance, "The widespread availability of consumer generated media (CGM) such as blogs, message boards, and news articles post great opportunities as well as risks to today’s enterprises." As of 2009 companies have already been applying this realization. The complexity issue is still relevant even when narrowing the search space to a single source of information. The challenge that exists after the search space is established is to locate the relevant data. After the relevant data is established it can then be assessed for sentiment. These two stages are commonly referred to as subjectivity classification and sentiment classification. "Subjectivity classification is a task to investigate whether a paragraph presents the opinion of its author or reports facts. Subjectivity classification can prevent the polarity [i.e. sentiment] classifier from considering irrelevant or even potentially misleading text" as suggested by Tang et al. [67]. Sentiment classification has some variation among designers of each approach but ultimately serves the same abstract purpose. In 2010, Cai et al. [68] suggests "Sentiment analysis traditionally emphasizes on classification of web comments into positive, neutral, and negative categories”. There are several variations of this tradition. A more common trend in recent research is to get more specific in defining the sentiment spectrum. In 2009, Tang et al. [67] mentions "Sentiment classification includes two kinds of

classification forms, i.e., binary sentiment classification and multi-class sentiment classification". This multi-class sentiment approach will likely be the standard of the future. Human emotion spans a much more complicated spectrum than the simple black and white notions of positive and negative. Human beings have the strange capability to love and to hate something at the same time. The user could portray negative and positive sentiments on the same product. This is easy for humans to decipher but much more complicated for a machine.

A few different approaches have been developed to create more accurate results. General polarity-based sentiment classification is a great step forward from the previous contextual only approaches. Cai et al. [68] mentions that "such analysis is useful, but it lacks insights on the drivers behind the sentiments." In 2010, Qiu [69] developed an idea titled "Dissatisfaction-oriented Advertising Sentiment Analysis" or DASA that combines traditional sentiment analysis with basic keyword matching. In this approach, the software detects the negative sentiment of certain products. In 2004, Kim and Hovy [70] mentions in their research that choosing an accurate sentiment analyzer tool is challenging while processing unstructured texts such as tweets. Bollen, Mao and Zeng [11] in their research extensively mentioned about understanding and analyzing unstructured text is becoming an increasingly popular field and includes a wide spectrum of problems such as sentiment analysis, key phrase extraction, topic modeling/extraction, aspect extraction and more. They discuss a simple approach to do lexicon-based analysis on words or phrases that impart negative or positive sentiment to a sentence the words *"bad"*, *"not good"* would belong to the lexicon of negative words, while *"good"*, *"great"* would belong to the lexicon of positive words. But this meant such lexicons must be manually curated, and even then, they were not always accurate. The phrases such as "not bad" which imparts a positive sentiment is hard to detect with simple lexicon-based analysis.

Massa et al. [71] presented a model that uses a mix of unsupervised and supervised techniques to learn word vectors capturing semantic term-document information as well as rich sentiment content. This model capture both semantic and sentiment similarities among words. Authors evaluate this model with document level and sentence level categorization task in the domain of online movie reviews.

Pang and Lee [29] in 2004 proposed a novel machine learning method that applies text categorization techniques to adjust to just the subjective portion of the document, which is in following process: (1) label the sentences in the document as either subjective or objective, discarding the latter; and then (2) Apply a standard machine learning classifier to the resulting extract.

In 2016, Chatterjee and Perrizo [66] focused on running sentiment analysis on the pulled tweets of selected investors to identify which of the stock ticker symbols have positive, neutral and negative sentiment score. They used a more robust approach using machine learning to train models that detect sentiment. For training the system a large dataset of text records that was already labeled with sentiment for each record was first obtained. The first step was to tokenize the input text into individual words, then apply stemming. Next, they constructed features from these words; these features are used to train a classifier. Upon completion of the training process, the classifier could then be used to predict the sentiment of any new piece of text. The crux of their research involved gathering the sentiment score accurately using a properly trained sentiment analysis tool. They have extensively trained and used the Microsoft Azure's sentiment analyzer tool on the stock related tweet data and have shown in their research how the performance of this tool outperforms the other commercial tools like the Stanford NLP sentiment analysis engine.

2.4. Stock Market Predictions

Stock price trend prediction is an active research area, as more accurate predictions are directly related to more returns in stocks. Therefore, in recent years, significant efforts have been put into developing models that can predict for future trend of a specific stock or overall market. Most of the existing techniques make use of the technical indicators. Some of the researchers showed that there is a strong relationship between news article about a company and its stock prices fluctuations. In the first formal theoretical study of prediction markets in 1992, Forsythe et al. [72] explored why individuals would spend time trading in such a market. Specifically, they listed five motivations for traders to participate in a political stock market experiment, which were (1) entertainment, (2) expected differences in information (confidence in their knowledge about the political event relative to other traders), (3) expected differences in information-processing ability (confidence in their ability to interpret news relative to other traders), (4) expected differences in their talents as traders, and (5) risk-seeking behavior. Forsythe et al. [72] expected these differences to attract a diverse group of experimental subjects and were able to confirm this belief when analyzing actual political stock market participants' demographic characteristics, political and ideological preferences, investments, and earnings. In the context of prediction markets, another issue of considerable practical importance (originally identified by Manski in 2004 [73]) is under which conditions prediction market prices reflect the true aggregate beliefs of the individual traders. To explore this issue, in 2006 Wolfers and Zitzewitz [74] proposed two simple models based on a log utility function, which lead to an equilibrium price in the market that is equal to the mean belief of traders. In 2004, Wolfers and Zitzewitz [82] also provided encouraging testimony of the ability of prediction markets to forecast uncertain future events. They found that "[...] simple market designs can elicit expected means or probabilities, more complex

markets can elicit variances, and contingent markets can be used to elicit the market's expectations of covariances and correlations [...]". In 2003, Berg et al. [75] used the Iowa Electronic Market's prediction of the outcomes of the 1988, 1992, 1996 and 2000 U.S. presidential elections to provide the first study of the long-run predictive power of forecasting markets, finding that their markets gave accurate forecasts at both short and long horizons (single day vs. weeks and months). In another study on the predictive power of prediction markets, in 2004, Tetlock [76] used data from tradesports.com, an online market which at that time allowed wagers on both sports events and financial market data. He showed that financial prediction markets can be surprisingly efficient with relatively low numbers of market participants. In contrast to the studies discussed so far, in 1996, Ortner [77] reported results from prediction markets run on election outcomes in Austria, where markets showed clear signs of manipulation and did not reliably provide forecasts of higher quality than polling organizations. Rather, the market's results in his experiment had been deliberately and successfully manipulated by a minority of traders to deviate from the market's earlier consensus opinion, at the same time influencing the prices of related markets. In 2003, Chen et al. [78] also deviated from the bulk of the prediction market literature, albeit in an entirely different way. While most studies reported on markets employing standard double auctions, in their experiment they performed a nonlinear aggregation of individuals' predictions based on said individuals' skills and risk attitudes, as determined in previous prediction rounds in the same market. The results from such a "weighted" prediction outperformed both the simple market and the best of the individuals. Overall, the diverse topics of studies on prediction markets and their heterogeneous findings underline the novelty of the field. While not specifically focusing on prediction markets, this study nonetheless offers new evidence on markets' ability to process information and harmonize expectations.

Following is discussion on previous research on sentiment analysis of text data and different classification techniques. In 2012, Nagar and Hahsler [79] in their research presented an automated text mining-based approach to aggregate news stories from various sources and create a News Corpus. The Corpus is filtered down to relevant sentences and analyzed using Natural Language Processing (NLP) techniques. A sentiment metric, called NewsSentiment, utilizing the count of positive and negative polarity words is proposed as a measure of the sentiment of the overall news corpus. They have used various open source packages and tools to develop the news collection and aggregation engine as well as the sentiment evaluation engine. They also state that the time variation of NewsSentiment shows a very strong correlation with the actual stock price movement. In 2011, Yu et al. [80] present a text mining-based framework to determine the sentiment of news articles and illustrate its impact on energy demand. News sentiment is quantified and then presented as a time series and compared with fluctuations in energy demand and prices. In 2011, J. Bean [83] uses keyword tagging on Twitter feeds about airlines satisfaction to score them for polarity and sentiment. This can provide a quick idea of the sentiment prevailing about airlines and their customer satisfaction ratings. In 2015, Shynkevich et al. [81] in their research studies show, how the results of financial forecasting can be improved when news articles with different levels of relevance to the target stock are used simultaneously. They used multiple kernels learning technique for partitioning the information which is extracted from different five categories of news articles based on sectors, sub-sectors, industries etc. News articles are divided into the five categories of relevance to a targeted stock, its sub industry, industry, group industry and sector while separate kernels are employed to analyze each one. The experimental results show that the simultaneous usage of five news categories improves the prediction performance in comparison with methods based on a lower number of news categories.

In 2016, Chatterjee and Perrizo [66] in their research show that the sentiment trend of key investors does bears relation with the actual stock movement in the market. With the Microsoft Azure Sentiment Analyzer running on the pulled tweets we can have the sentiment score generated on the pulled tweets for every investor on particular ticker symbol and can observe the trend of the particular ticker over a period of time. A single sentiment score is assigned to the complete tweet. The Azure sentiment analyzer for every tweet generates a score between 0-1 where sentiment score > 0.7 denotes high positive sentiment while < 0.4 denotes low negative sentiment. The sentiment score range between 0.4-0.7 denotes the neutral range. In their research, they choose various time spans on which the sentiment score has been generated from the extracted tweets and observe how the sentiment line varies. The sentiment analyzer generates a sentiment score trend line over a time span and based on this line strategic investment decisions can be made.

CHAPTER 3. TWITTER, INVESTOR SENTIMENTS AND SENTIMENT ANALYSIS

Microblogging today has become an extremely popular communication tool among Internet users. Millions of messages are appearing daily in popular web-sites that provide services for microblogging such as Twitter, Tumblr, Facebook. Authors of those messages write about their life, share opinions on variety of topics and discuss current issues. Because of a free format of messages and an easy accessibility of microblogging platforms, Internet users tend to shift from traditional communication tools (such as traditional blogs or mailing lists) to microblogging services. As more and more users post about products and services they use, or express their political and religious views, microblogging web-sites become valuable sources of opinions and sentiments of people.

Twitter is one of the most important social media platforms in today's world providing the unique ability for a user to connect with almost anyone else in the world. The platform supports 34 languages and has close to 317 million active users. Every second, on average, around 6,000 tweets are tweeted on Twitter, which corresponds to over 350,000 tweets sent per minute, 500 million tweets per day and around 200 billion tweets per year [31]. Investors are day by day using the Twitter platform to cite their opinions on particular ticker symbols and share their market focused posts and updates. With so much information in the platform it is difficult to find the information a particular user just needs to reference to make an investment decision [11]. In the realm of stocks, it is important to understand which ticker symbols to follow on which investment decisions can be made. So, it is important to follow investors and find what the common ticker symbols and trends they are discussing on. Sentiment is generally defined as a thought view or attitude and enables large scale understanding and clarity regarding the feelings of a group of people (in this context investors) on a given subject (in this context stocks) [11, 12, 13, 14]. Once

the sentiments and trends of their discussion are analyzed then a user can make a constructive business decision [24, 26]. The first section of this chapter highlights the aspects a programmer needs to take into consideration while programming around Twitter platform and the second section highlights Twitter as a corpus for investor sentiments how it affects the cross-section of stock returns.

3.1. Limitations of the Twitter Platform

This section explains in detail that how tweets can be programmatically retrieved. Twitter has a REST API that allows users to search for tweets, users, timelines, or even post new messages. We use Tweetinvi which is a C# based .net API which is been used to access the Twitter API. A Twitter developer account needs to be set up first with the necessary credentials to query the API using Tweetinvi. The project is layered to keep the twitter interactions, file management, application logic and algorithms separate. The following code snippet first sets the twitter credentials for querying against the API and then from a list of users from a user list pulls the tweets and includes the logic to recalculate the timespan between the tweets to optimize the API pull requests from Twitter.

```
namespace TwitterScraper
{
    0 references
    class Program
    {
        0 references
        static void Main(string[] args)
        {
            //Credentials - Get your own credentials from http://dev.twitter.com
            Twitter.SetCredentials(accessToken: "11111-11111-111111", accessTokenSectet: "11111-11111-111111",
                                   consumerKey: "11111-11111-111111", consumerSecret: "11111-11111-111111");

            Console.WriteLine("Credentials ready!");

            //Initial Variables
            var startDate = DateTime.Now;

            //Start program
            while (true)
            {
                //Read Twitter Users List
                var usernames = FileManagement.GetUsernames();
                var nextUpdateTime = FileManagement.GetNextUpdateTime();

                foreach (var userName in usernames)
                {
                    //Directory to save files
                    var file = FileManagement.GetFilename(userName);
                    int tweetCount = 0;

                    //Check the time for the next download
                    if (nextUpdateTime[userName] > DateTime.Now)
                    {
                        continue;
                    }

                    //Recalculate time span between tweets
                    var tweets = FileManagement.GetTweets(file);
                    var timeSpan = Twitter.GetAverageTimeSpan(tweets);
                    nextUpdateTime[userName] += timeSpan;

                    //Get the tweets for this user
                    var tweetList = Twitter.GetTimeline(userName, tweets, ref tweetCount);
                    Twitter.Statistics(tweetCount, userName);

                    //Encode each tweet and add them to a list
                    var encodedTweetList = FileManagement.Serializer(tweetList);

                    //Open & Write to file only if there are new tweets
                    FileManagement.Writer(encodedTweetList, file);
                    FileManagement.Logger(encodedTweetList, userName);
                }
            }
        }
```

Fig. 1. Code snippet to pull tweets for a collection of usernames

Based on the sector we would be interested in following; we first create a list of usernames (tweet ID's) from that particular sector whose tweets we would like to analyze. The table below shows the combination of UserId, UserName, LastTweetId and IsEnabled attributes which we track. The IsEnabled attribute denotes whether the user still has an active twitter account or not.

Table 1. UserID, UserName, LastTweetID, IsEnabled attributes

UserID	UserName	LastTweetID	IsEnabled
1	alexforrest	853160224328486913	1
2	fionawalsh	853681056503009280	1
5	katedevlin	849538565168140288	1
6	kathrynhopkins	851868730946662400	1
7	kenkaufman	848643257349406721	1
8	10000words	825104404420317185	1
9	140elect	600454114473017345	1

For our analysis, we identified and utilized a list of 3000 financial and news symbol that we took from twitter lists and search results. We also estimated when it was the best time to update the users tweet based on their tweeting frequency. The following snippet of code shows how we are getting the ID of the latest saved tweet, the users average posting frequency and adding the timestamp to the pulled tweets.

```
/// Get the ID of the latest saved tweet
/// </summary>
1 reference
public static long LastSavedTweetID(List<Tweet> getTweets)
{
    var lastLine = getTweets.First();
    long lastTweetID = lastLine.ID;
    return lastTweetID;
}

/// <summary>
/// User's average posting frecuency
/// </summary>
1 reference
public static TimeSpan GetAverageTimeSpan(List<Tweet> tweets)
{
    if (tweets.Count == 0)
    {
        return TimeSpan.FromSeconds(500);
    }
    else
    {
        List<DateTime> dates = new List<DateTime>();
        foreach (var line in tweets)
        {
            dates.Add(line.Time);
        }
        var difference = dates.Max().Subtract(dates.Min());
        var averageTimes = TimeSpan.FromMilliseconds(difference.TotalMilliseconds / (dates.Count()));
        return averageTimes;
    }
}

/// <summary>
/// Add Time Stamps to list
/// </summary>
1 reference
private static void AddTimeStamps(List<DateTime> timeStamps)
{
    if (timeStamps.Count < 300)
    {
        timeStamps.Add(DateTime.Now);
    }
    else
    {
        timeStamps.RemoveAt(0);
        timeStamps.Add(DateTime.Now);
    }
}
```

Fig. 2. Code snippet to calculate average posting frequency and latest saved tweets

The freshly downloaded tweets are serialized to JSON by JSON.net and then written to an .xml-based file one per twitter username. The program is designed to download the maximum historical tweets possible per user and then rechecking the accounts for new tweets. Twitter limits the API requests to 300 requests per 15 minutes and allows access to maximum of 3200 historical tweets. Considering the financial sector, we would not be missing much information as the historical tweets will not have much significance as compared to the present state of the markets.

Additionally, each request to the platform can download a maximum of 200 tweets at a time. To maximize the productive use of the requests, calculating the average time span from the user's latest tweets and setting the time the program to recheck for new tweets is necessary. At any given instance, we scan for the latest tweets which are made by all the users and have a timestamp associated with those pulled tweets. For every 5 minutes from the latest pulled timestamp, we then check if there are any new updates which are made by the users, and we then download the tweets which are made in those 5 minutes. The code snippet shows every time when the API is ready for a receiving a new request, the statistics of the current download with API status and adding a timestamp to the list.

```
/// <summary>
/// Check if API is ready for new request
/// </summary>
2 references
private static int WaitForAPIReady()
{
    int count = 0;
    do
    {
        DateTime currentTime = DateTime.Now;
        currentTime = currentTime.AddMinutes(-15);

        count = (from time in timeStamps
                 where time > currentTime
                 select time).Count();
        if (count > 290)
        {
            Console.WriteLine(" Twitter downloading limit reached. Waiting...");
            Thread.Sleep(50000);
        }
    } while (count > 290);
    return (300 - count);
}

/// <summary>
/// Statistics of current download and API status
/// </summary>
1 reference
public static void Statistics(int tweetsDownloaded, string userName)
{
    int requestsLeft = WaitForAPIReady();
    Console.WriteLine(String.Format(" {2} |  Requests left: {0}  |  Tweets Downloaded: {1}\n", requestsLeft, tweetsDownloaded, userName));
}

/// <summary>
/// Add Time Stamps to list
/// </summary>
1 reference
private static void AddTimeStamps(List<DateTime> timeStamps)
{
    if (timeStamps.Count < 300)
    {
        timeStamps.Add(DateTime.Now);
    }
    else
    {
        timeStamps.RemoveAt(0);
        timeStamps.Add(DateTime.Now);
    }
}
```

Fig. 3. Code snippet to check API status and setting the timestamp

While downloading the new tweets, we check for the already downloaded tweets of the user and if there are historical tweets then we will capture the updates only. Each tweet has its own format and contains a lot of information (ID, Text, Time, Retweets, Favorites, User, and Followers).

3.2. Investor Sentiments and Tweet Counts

In our research, we are using information from 3000 very well-known twitter users ranging from individual stock investors, financial advisors to news channels between January 2015 and December 2017. The selection process of choosing these 3000 investors has been guided through multiple decision criteria some of them being – the top CNN list of analysts from the last 5 years, Wall Street Journals accredited financial analysts, very well-known investors such as Warren Buffet and Jim Cramer who have a proven history of consistently beaten the market with their predictions and stock advisors like Motley Fool who have a proven recognition. The selection process of individual financial advisors has been governed by the amount of wealth management which households have entrusted them with. Online news articles from Barrons, helped pick some of the most notable financial advisors like Colleen O'Callaghan managing $2.8 billion from 100 households, Thomas Moran managing $3.1 billion from 1250 clients, Michael Klein managing $8.7 billion for 335 clients as examples of users who have a proven credibility. The process of the selection of these investors is a manual effort but certain advanced decision criterions could also be placed.

Investors can safely be assumed to be sentiment driven. The top down approach [7, 8] focuses on the measurement of reduced form, aggregate sentiment and traces its effects to market returns and individual stocks. The approach is based on two broad undisputable assumptions of behavioral finance— sentiment and the limits to arbitrage—to explain which stocks are likely to

be most affected by sentiment, rather than simply pointing out that the level of stock prices in the aggregate depends on sentiment. Stocks of low capitalization, younger, unprofitable, high volatility, non-dividend paying, growth companies, or stocks of firms in financial distress, are likely to be disproportionately sensitive to broad waves of investor sentiment [8, 10].

We are pulling the tweets from these investors in real time and parsing the tweets for the ticker symbol. Each ticker symbol precedes the "$" symbol and follows a "$<Ticker>" pattern within the tweet text. The ticker symbol with the highest frequency of occurrence is the most discussed stock amongst these investors [40]. This insight can be helpful especially when a user needs to know which stocks to invest in primarily and does not have much knowledge on the markets. Table 2 shows the ticker symbols which are frequently discussed amongst the investors with their respective tweet counts.

Table 2. Top 10 Ticker symbols and tweet counts between 2015/01/05 and 2017/12/05.

Ticker Symbol	Tweet Counts
$SPY	15190
$AAPL	9232
$SPX	8866
$VIX	5724
$AMZN	5542
$FB	5190
$QQQ	5018
$TSLA	4329
$GLD	3165
$NFLX	3155

Also, in our analysis, we focus on investors being sentiment driven on particular ticker symbols based on their tweet counts. The more is the heterogeneity of the population of investors tweeting on *s*, signifies *s* is being discussed by a wider population range. If a stock is discussed by

a wider group of users, then the volume of the stocks traded for a given day is seen to significantly increase. Studies from previous research [84] show the number of trades (stock volume) was correlated with the number of tweets which were discussed by a broader sample of investors. In this research, we have tried to diversify the selection process for choosing the stocks and have tried to choose stocks of companies from various sectors as shown in table 3.

Table 3. Companies with ticker symbols selected across different sectors.

Sectors	Company	Symbol
Basic Materials Industries: Chemicals, Energy, Metals & Mining	Exxon Mobile Corp	XOM
	Schlumberger Ltd	SLB
Consumer Goods	Apple Corp.	AAPL
	Coca-Cola Co	KO
Financial	Wells Fargo	WFC
	Citigroup	C
Health Care	Gilead Sciences	GILD
	Pfizer Inc	PFE
Sectors	Company	Symbol
Industrial Goods	3M Co.	MMM
	Caterpillar Inc	CAT
Services	Amazon Inc.	AMZN
	Netflix	NFLX
	Facebook	FB
	McDonalds Corp.	MCD
Technology	Microsoft Corp.	MSFT
	Alphabet Inc.	GOOG
	Alibaba Group	BABA
Utilities	Duke Energy	DUK
	Exelon Corp.	EXC

From this list of stocks across the different sectors shown in table 3, we try to find the stocks which are mostly been discussed by the 3000 investors we are tracking, and the results are shown in table 4. In this research we will focus on these identified top 10 company and ticker symbols.

Table 4. Top 10 Company Ticker symbols with respective tweet counts.

Ticker Symbol	Tweet Counts
$AAPL	9232
$AMZN	5542
$FB	5190
$NFLX	3155
$GOOG	1930
$BABA	1616
$MSFT	1383
$GILD	1273
$C	848
$MCD	806

3.3. Sentiment Analysis

The main objective of this research is to run the sentiment analysis on the set of tweets of selected investors so that we can identify which of the ticker symbols have positive, neutral or negative sentiment scores. Choosing an accurate sentiment analyzer tool can be challenging while processing unstructured texts such as tweets. Understanding and analyzing unstructured text is an increasingly popular field and includes a wide spectrum of problems such as sentiment analysis [5, 6], key phrase extraction, topic modeling/extraction, aspect extraction and more. Sentiment Analysis involves several key challenges. One simple approach is to do lexicon-based analysis on

words or phrases that impart negative or positive sentiment to a sentence the words “bad”, “not good” would belong to the lexicon of negative words, while “good”, “great” would belong to the lexicon of positive words. But this means such lexicons must be manually curated, and even then, they are not always accurate [36]. The phrases such as “not bad” which imparts a positive sentiment is hard to detect with simple lexicon-based analysis.

3.3.1. Azure Sentiment Analyzer

A more robust approach is to use machine learning to train models that detect sentiment. For training the system a large dataset of text records that was already labeled with sentiment for each record was first obtained. The first step is to tokenize the input text into individual words, then apply stemming. Next, we constructed features from these words; these features are used to train a classifier. Upon completion of the training process, the classifier can be used to predict the sentiment of any new piece of text. It is important to construct meaningful features for the classifier [50, 51], and our list of features includes several from state-of-the-art research:

- **N-grams** denote all occurrences of *n* consecutive words in the input text. The precise value of *n* may vary across scenarios, but it’s common to pick *n=2* or *n=3*. With *n=2*, for the text *“the quick brown fox”*, the following n-grams would be generated – *[“the quick”, “quick brown”, “brown fox”]*

- **Part-of-speech tagging** is the process of assigning a part-of-speech to each word in the input text. We also compute features based on the presence of emoticons, punctuation and letter case (upper or lower)

- **Word embedding’s** are a recent development in natural language processing, where words or phrases that are syntactically similar are mapped closer together, e.g. in such a mapping, the term *cat* would be mapped closer to the term *dog*, than to the term *car*, since both

dogs and cats are animals. Neural networks are a popular choice for constructing such a mapping. For sentiment analysis, we employ neural networks that encode the associated sentiment information as well. The layers of the neural network are then used as features for the classifier. So, the crux of the research involves gathering the sentiment score accurately using a properly trained sentiment analysis tool. In this research, we have used Microsoft Azure Sentiment Analyzer to run the sentiment analysis on the pulled tweets.

3.3.2. Comparing Azure Sentiment Analyzer with other NLP Tools

We evaluated the performance of the classifier [36] against two external services the Stanford NLP Sentiment Analysis engine (using its pre-trained sentiment model), and a popular commercial tool. Here are the comparative benchmarks - On datasets comprising tweets, Azure ML Text Analytics was 10-20% better at identifying tweets with positive vs negative sentiment. The data sets we used were from the Sentiment140 and CrowdScale datasets. The Sentiment140 dataset comprises approximately 1.6 million automatically annotated tweets. The tweets were collected by using the Twitter Search API and keyword search. During automatic annotation, any tweet with positive emoticons, like :), were assumed to bear positive sentiment, and tweets with negative emoticons, like :(, were supposed to bear negative polarity. Tweets containing both positive and negative emoticons were removed. Additional information about this data and the automatic annotation process can be found in the technical report written by Go et al. [87]. Each instance in the data set has 6 fields:

- sentiment_label - the polarity of the tweet (0 = negative, 2 = neutral, 4 = positive)
- tweet_id - the id of the tweet
- time_stamp - the date of the tweet (Sat May 16 23:58:44 UTC 2009)
- target - the query. If there is no query, then this value is NO_QUERY

- user_id - the user who posted the tweet
- tweet_text - the text of the tweet

However, for the experiment on Sentiment140 dataset we have used only two fields that are required for training - the sentiment label and the tweet text.

Table 5. Azure ML Text Analytics result on Sentiment140 dataset.

True Positive	False Negative	Accuracy	Precision	AUC
129808	30192	0.79	0.77	.86
False Positive	True Negative	Recall	F1 Score	
37074	122926	.81	.79	

CrowdScale dataset is another sentiment analysis judgement dataset. The tweets in the dataset are from the weather domain. Each tweet was evaluated by at least 5 raters. The possible answers are: “Negative”, “Neutral”; the author is just sharing information, “Positive”, “Tweet not related to weather condition” and “I can’t tell”. The tweets from the test set of the CrowdScale dataset were evaluated from Azure ML, Stanford and a commercial available tool and the responses from the experiment is in the table 6.

Table 6. Responses from the different Sentiment Analysis Tools on CrowdScale dataset.

Tool	Identified Positive Response	Identified Negative Response	Identified Neutral Response
Azure ML	2,825	3,052	3,937
Stanford	2,492	2,691	3,472
Commercial Tool	2,363	2,552	3,292

We analyzed sentiment on a dataset of TripAdvisor reviews as well. Three equally experienced annotators provided sentence-level annotations of a subset of 500 randomly selected reviews from the publicly available TripAdvisor dataset [86]. The full TripAdvisor dataset consists of 235,793 hotel reviews crawled over a period of a month. In addition to the review text, each review comes with a hotel identifier, an overall rating and optional aspect-specific ratings for the following seven aspects: Rooms, Cleanliness, Value, Service, Location, Checkin, and Business. All review-level ratings are on a discrete ordinal scale from 1 to 5 (with -1 indicating that an aspect-specific rating was not provided by the reviewer). The annotation distinguishes between Positive, Negative and Neutral/Mixed opinions. The Neutral/Mixed label is assigned to opinions that are about an aspect without expressing a polarized opinion, and to opinions of contrasting polarities, such as "The room was average size" (neutral) and "Pricey but worth it!" (mixed). The annotations also distinguish between explicit and implicit opinion expressions, that is, between expressions that refer directly to an aspect and expressions that refer indirectly to an aspect by referring to some other property/entity that is related to the aspect. For example, "Fine rooms" is an explicitly expressed positive opinion concerning the Rooms aspect, while "We had great views over the East River" is an implicitly expressed positive opinion concerning the Location aspect, and "All doors seem to have to slam to close" is an implicit negative opinion concerning the Rooms aspect. The final dataset consists of 369 unique reviews partitioned into a training set (258 reviews, 70% of the total) and a test set (111 reviews, 30% of the total). The data was split by selecting reviews for each subset in an interleaved fashion, so that each subset constitutes a minimally biased sample both with respect to the full dataset and with respect to annotator experience. On the annotated training dataset of 111 reviews, 71 were positive and 40 were negative and here is a comparison of the results from the sentiment tools as well as the F1 score for the positive and negative reviews.

Table 7. Responses from the different sentiment tools on TripAdvisor dataset

Tool	Identified Positive Reviews	Identified Negative Reviews
Azure ML	56	25
Stanford	46	21
Commercial Tool	44	18

The Azure Machine Learning Text Analytics [36] outperforms other offerings on short as well as long forms of text for the sentiment analysis task. In this research we will run Azure Sentiment Analyzer and compute the sentiment score on every pulled tweet from the 3000 investors. For any given day, we parse the tweets which contain references to the top 10 company ticker symbols from table 4 and compute the daily average sentiment score of the stock for that day. The daily average sentiment score (AvgSentiment) is computed for every day on every stock by taking the mean of the sentiment scores from the number of pulled tweets for a certain day. If different stock symbols are mentioned in the same tweet, the same sentiment score from Microsoft Azure Sentiment Analyzer would be used for each ticker symbol. The Azure sentiment analyzer for every tweet generates a score between 0-1 where sentiment score > 0.7 denotes high positive sentiment while < 0.4 denotes low negative sentiment. The sentiment score range between 0.4-0.7 denotes the neutral range. If for a period, there are no sentiments which get generated (if there are no investors tweeting) for a particular stock we preserve the Sentiment score which is preserved from the last day the sentiments were computed. The snapshot of the pulled data for one of the companies “Apple Inc.” from the consumer goods sector is shown in the table 8.

Table 8. Apple Inc. stock data with sentiment score between 2015-02-02 till 2015-02-12.

Ticker	Trade Date	Open Price	Close Price	High Price	Low Price	Avg Sentiment	Trade Volume
AAPL	2015-02-02	118.05	118.63	119.17	116.08	0.75	62739100
AAPL	2015-02-03	118.50	118.65	119.09	117.61	0.54	51915700
AAPL	2015-02-04	118.50	119.56	120.51	118.31	.61	70149700
AAPL	2015-02-05	120.02	119.94	120.23	119.25	.63	42246200
AAPL	2015-02-06	120.02	118.93	120.25	118.45	.79	43706600
AAPL	2015-02-09	118.55	119.72	119.84	118.43	.68	38889800
AAPL	2015-02-10	120.17	122.02	122.15	120.16	.58	62008500
AAPL	2015-02-11	122.77	124.88	124.92	122.5	.61	73561800
AAPL	2015-02-12	126.06	126.46	127.48	125.57	.58	74474500

In the next chapter, we will discuss how we will use the sentiment score to predict the close price of the stock using different regression models.

CHAPTER 4. BUILDING PREDICTION MODELS USING SENTIMENT SCORE

4.1. Introduction

Stock market prediction has been an active area of research for a long time. The Efficient Market Hypothesis (EMH) states that stock market prices are largely driven by new information and follow a random walk pattern [37, 39, 41]. Though this hypothesis is widely accepted by the research community as a central paradigm governing the markets in general, several people have attempted to extract patterns in the way stock markets behave and respond to external stimuli. In this dissertation, we test a hypothesis based on the premise of behavioral economics, that the emotions and moods of individuals affect their decision-making process, thus, leading to a direct correlation between "public sentiment" and "market sentiment" [42, 43, 44, 45]. We also conduct various experiments on different time periods using regression-based prediction models [15,16, 19, 20] and calculate the coefficient of determination of predicting the close price of the stock.

4.2. Literature Review

The work in this dissertation is based on the strategy from Bollen et al. [10,11,12,13] and Mittal et al. [14]. They also attempted to predict the behavior of the stock market by measuring the mood of people on Twitter. The authors considered the tweet data of all twitter users in 2008 and used the OpinionFinder and Google Profile of Mood States (GPOMS) algorithm to classify public sentiment into 6 categories, namely, Calm, Alert, Sure, Vital, Kind and Happy. They cross validated the resulting mood time series by comparing its ability to detect the public's response to the presidential elections and Thanksgiving Day in 2008. They also used causality analysis to investigate the hypothesis that public mood states, as measured by the OpinionFinder and GPOMS mood time series, are predictive of changes in DJIA close values. The authors used Self Organizing Fuzzy Neural Networks to predict DJIA values using previous values. Their results show a

remarkable accuracy of nearly 87% in predicting the up and down changes in the close values of Dow Jones Industrial Index (DJIA).

In this dissertation, we are considering 3000 investors whom we track and rather than classifying the sentiments into categories we generate a sentiment score between 0 and 1 for every pulled tweet. Once the daily sentiment scores are computed we combine this information with the fundamentals of the stock (Open Price, Close Price, Volume, High Price, Low Price) and use conventional ARIMA based time series models and regression-based prediction models [47, 48, 49] to calculate the coefficient of determination. But before we delve into building different prediction models, we first use Granger Causality Analysis [32, 33, 34] to understand the lag period of sentiments to have a considerable effect on the close price of the stock.

4.3. Granger Causality

Granger Causality analysis [32, 35, 38] finds how much predictive information one signal has about another over a given lag period. The p-value measures the statistical significance of our result i.e. how likely we could obtain the causality value by random chance; therefore, lower the p-value, higher the predictive ability. The sentiment score for a given day is measured for a span of 24 hours while the close price of the stock is determined based on the time the market is open (NYSE: 9:30-16:00 EST). So, the sentiment score can be carried forward to affect the open price of the stock for the next day. We performed the Granger Causality test for studying the effect of sentiment score on the close Price of the stock, for studying the effect of close price on the sentiment score of the stock, the effect of open price on the sentiment Score of the stock and for studying the effect of sentiment score on the open price of the stock for lag periods of 1, 2, 3, 7 and 10 days.

Table 9. p-values obtained using Granger causality analysis with different lag periods.

Lags (days)	Symbol	ClosePrice ~ AvgSentiment	AvgSentiment ~ ClosePrice	OpenPrice ~ AvgSentiment	AvgSentiment ~ OpenPrice
1	AAPL	.66	.005*	.25	.004*
2		.62	.002*	.52	.002*
3		.77	.017*	.62	.017*
7		.93	.11	.88	.08*
10		.96	.21	.96	.16
1	MSFT	.82	.55	.94	.51
2		.29	.83	.82	.69
3		.14	.81	.65	.86
7		.25	.87	.53	.76
10		.30	.75	.70	.71
1	GOOG	.45	.02*	.93	.03*
2		.71	.06*	.37	.07*
3		.37	.09*	.21	.09*
7		.75	.23	.18	.16
10		.71	.23	.20	.26

The results shown in the table 10 suggests sentiment score of the stock for a given day is seen to have effect from the Close Price of the stock and the Open Price of the stock for the following day in most cases have an effect from the sentiment score of the previous day.

4.4. Coefficient of Determination

The *coefficient of determination* (R^2) is a measure that allows us to determine how certain one can be in making predictions from a certain model/graph. It is the ratio of the explained variation to the total variation. It is also a measure how well the regression line

represents the data. If the regression line passes exactly through every point on the scatter plot, it would be able to explain the variation. The further the line is away from the points, the less it can explain. R^2 is a statistical factor that will give some information about the goodness of fit of a model. In regression, the R^2 coefficient of determination is a statistical measure of how well the regression line approximates the real data points. An R^2 of 1 indicates that the regression line perfectly fits the data. R^2 is often interpreted as the proportion of response variation "explained" by the regressors in the model. An interior value such as $R^2 = 0.7$ may be interpreted as follows - Seventy percent of the variance in the response variable can be explained by the explanatory variables. The remaining thirty percent can be attributed to unknown, lurking variables or inherent variability.

In case of a single regressor as in this dissertation, fitted by least squares, R^2 is the square of the Pearson product-moment correlation coefficient relating the regressor and the response variable. More generally, R^2 is the square of the correlation between the constructed predictor and the response variable. With more than one regressor, the R^2 can be referred to as the coefficient of multiple determination.

4.5. Regression Models

Regression algorithms are algorithms that fit the values of a real function for a single instance of data. Regression algorithms can incorporate input from multiple features [21, 22, 23], by determining the contribution of each feature of the data to the regression function. Once the regression algorithm has trained a function based on already labeled data, the function can be used to predict the label of a new (unlabeled) instance. In this section, we discuss about several experiments using Bayesian Linear Regression for predicting the close price of the stock. We first preprocess the daily stock data with the sentiment scores as shown in table 8 and convert it into a

matrix-based format. This means that the stock price on the n^{th} day will be affected mostly by the close price of the stock and the sentiment score for the previous n-1 days. From the Granger causality analysis, we have seen that for the sentiments to have an effect on the close price of the stock we should back at no longer than 7 days. We consider all the stocks from table 4 for our analysis and we perform three different experiments – i) Choosing 5-day trading day period. ii) Choosing 7-day period including Saturday and Sunday when the markets are closed but we have sentiment data from the tweets and iii) Choosing 7-day period where Saturday and Sunday are excluded from the analysis since the markets are closed. We calculate the *coefficient of determination* (R^2) for each of these experiments and evaluate the optimum time to select.

4.5.1. Training the Model

Training a classification or regression model is a kind of *supervised machine learning.* We provide a dataset that contains historical stock and sentiment data from which the model can learn patterns. The data should contain both the outcome we are trying to predict, and related factors (variables). The machine learning model uses the data to extract statistical patterns and build a model. The data set is split into 80-20 where the initial 80% of the dataset is first used to train the model and once trained it is tested on the remaining 20%. We will now discuss the different regression models we are focusing on in this dissertation using Azure Machine Learning Studio [54]. In this dissertation, we will be applying various regression model modules from the Azure Machine Learning Studio on the matrix converted pre-processed data sets and compare the statistical metrics - Mean Absolute Error, Root Mean Squared Error, Relative Absolute Error, Relative Squared Error and the Coefficient of Determination.

4.5.2. Bayesian Linear Regression

The Bayesian Linear Regression is a regression model based on Bayesian statistics. Bayesian approach uses linear regression supplemented by additional information in the form of a prior probability distribution. Prior information about the parameters is combined with a likelihood function to generate estimates for the parameters. In this dissertation, we use the Bayesian Linear Regression module [17, 18] to create a regression model based on Bayesian statistics. A classical treatment of regression [25] problem seeks a point estimate of the unknown parameter vector *w*. By contrast, in a Bayesian approach we characterize the uncertainty in w through a probability distribution *p(w)*. Observations of data points modify this distribution by Bayes theorem, with the effect of the data being mediated through the likelihood function. Specifically, we define a prior distribution *p(w)* which expresses our uncertainty in *w* taking account of all information aside from the data itself, and which, without loss of generality, can be written in the form $p(w|\alpha) \propto \exp\{-\alpha\Omega(w)\}$ where, α can again be regarded as a hyperparameter.

Once we have configured the model, we train the model using a tagged dataset and the Train Model module in the Azure Machine Learning Studio. The trained model is then used to make predictions.

4.5.2.1. Experiments to Select the Optimal Days

In this dissertation, we conduct several experiments to see which time period within a week should be considered and we run Bayesian Linear Regression based prediction models to predict the closing price of those stocks. From Granger Causality, we understand that news and twitter messages have considerable effect on the stock price and the effect can only be for a week. In the experiments we focus on the top 10 most discussed stocks amongst our chosen investors as shown earlier from table 4.

4.5.2.1.1. Experiment 1 – 5 Day Period

For this experiment, we have built a Bayesian Linear Regression model for 5 days period. The dataset only considers 5 working days data within a week for a given ticker name. The dataset has data for every stock from 01/01/2015 – 12/04/2017. The data set has been split into 80-20 training and test data respectively with the weekend sentiment score data points removed. Table 10 shows the performance matrix on the test dataset.

Table 10. Performance matrix on dataset for 5-day period.

Model	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Relative Squared Error	Coefficient of Determination
AAPL	1.35	1.79	0.29	0.08	0.92
AMZN	5.03	8.65	0.11	0.02	0.97
FB	0.85	1.15	0.20	0.04	0.95
NFLX	1.35	2.42	0.08	0.02	0.98
GOOG	6.13	10.47	0.28	0.01	0.90
BABA	1.10	1.60	0.24	0.06	0.93
MSFT	0.47	0.75	0.22	0.10	0.89
GILD	1.25	1.70	0.12	0.05	0.94
C	0.52	0.68	0.22	0.05	0.94
MCD	0.75	1.01	0.42	0.18	0.81

We can see the Coefficient of Determination for this model range from 0.81-0.98 for between our chosen ticker symbol and seven ticker symbols above 0.94 Coefficient of Determination.

4.5.2.1.2. Experiment 2 –7 Day Period (Excluding Weekends)

In this experiment, we have built Bayesian Linear Regression model excluding weekend data but for 7 days instead of 5 days in experiment 1. The dataset has data for every stock from 01/01/2015 – 12/04/2017 and does not include weekend sentiment scores. The data set has been split into 80-20 training and test data respectively. Table 11 shows the performance matrix on this test dataset.

Table 11. Performance matrix on dataset for 7-day period (excluding Saturday and Sunday).

Model	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Relative Squared Error	Coefficient of Determination
AAPL	1.36	1.80	0.29	0.08	0.91
AMZN	5.04	8.72	0.11	0.02	0.97
FB	0.85	1.15	0.21	0.04	0.95
NFLX	1.40	2.45	0.08	0.02	0.98
GOOG	6.13	10.49	0.28	0.10	0.89
BABA	1.09	1.59	0.24	0.06	0.93
MSFT	0.47	0.75	0.22	0.10	0.89
GILD	1.26	1.72	0.19	0.05	0.94
C	0.53	0.70	0.22	0.05	0.94
MCD	0.77	1.01	0.43	0.18	0.81

We can see the Coefficient of Determination for this model range from 0.81-0.98 for between our chosen ticker symbol and five ticker symbols above 0.94 Coefficient of Determination.

4.5.2.1.3. Experiment 3 – 7 Day Period (Including Weekends)

In this experiment, we have built Bayesian Linear Regression model for 7 days period including weekend data. Since we don't have close price of the stocks for weekend we are taking Friday's closing stock fundamental values as static values for Saturday and Sunday along with the daily average sentiment score for the stocks on Saturdays and Sundays. The dataset for this training experiment includes closed price with daily average sentiment score value of a given ticker name. The dataset has data for every stock from 01/01/2015 – 12/04/2017. The data set has been split into 80-20 training and test data respectively with the weekend sentiment score data points included. Table 12 shows the performance matrix on this test dataset.

Table 12. Performance matrix on dataset for 7-day period (including weekends).

Model	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Relative Squared Error	Coefficient of Determination
AAPL	1.01	1.51	0.22	0.05	0.94
AMZN	3.8	7.26	0.08	0.01	0.98
AAPL	1.01	1.51	0.22	0.05	0.94
AMZN	3.8	7.26	0.08	0.01	0.98
FB	0.66	0.98	0.16	0.03	0.96
NFLX	1.05	2.06	0.06	0.01	0.98
GOOG	4.76	8.84	0.21	0.06	0.93
BABA	0.82	1.34	0.18	0.04	0.95
MSFT	0.35	0.64	0.16	0.07	0.92
GILD	0.93	1.44	0.14	0.040	0.95
C	0.38	0.57	0.16	0.03	0.96
MCD	0.57	0.85	0.32	0.13	0.86

We try and validate how much of an effect weekends tweets have on the performance of our model. Our observation is, we can see improved performance of each ticker symbol when we include the weekend sentiment data for predicting price for a given day. We see the Coefficient of Determination for this model range from 0.92-0.98 for our chosen ticker symbols and nine ticker symbols above 0.92 Coefficient of Determination.

4.5.2.1.4. Conclusions from the Experiments

We are considering Coefficient of Determination as one of the key measure to evaluate and compare performance of each experiment. The resulting performance matrix of all experiments show that coefficient of Determination is optimal when we consider 7 days data set with weekend sentiment scores. For most of the tickers, our experiment shows descent improvement of coefficient of determination for the model that includes Saturday and Sunday's data. Table 13 shows the comparison of coefficient of determination between the different experiments.

Table 13. Coefficient of Determination between the experiments.

Model	Coefficient of Determination: 5-day period Excluding Weekend	Coefficient of Determination: 7-day period (excluding Saturday and Sunday)	Coefficient of Determination: 7-day period (including Saturday and Sunday)
AAPL	0.920673	0.919536	0.941159
AMZN	0.975499	0.975115	0.983221
FB	0.956013	0.9557	0.968316
NFLX	0.981877	0.981399	0.987234
GOOG	0.900253	0.899913	0.932627
BABA	0.932595	0.933515	0.951213
MSFT	0.894668	0.893938	0.92706
GILD	0.944528	0.943601	0.959782
C	0.94668	0.944525	0.963232
MCD	0.815396	0.811403	0.865635

It is evident that the model with 7 day period including Saturday and Sunday sentiment score is the most optimum model to use for prediction.

4.6. Predicting Stock Close Price

In this section, we will predict the close price of the stocks using the Bayesian Linear Regression model from the training experiment, trained till 12/04/2017 with the values of the sentiment scores and the close price of the stocks as input. We have created the training models for all 10 tickers with 7 days including (Saturday & Sunday) data and predict next two weeks of close price for the stocks and compare the results of our predictive experiment to the actual close price of the stock in Table 14.

Table 14. Predicted stock close price of tickers compared to their actual stock close price.

Date	MSFT		NFLX		AAPL	
	Predicted	Actual	Predicted	Actual	Predicted	Actual
12/5/2017	81.18	81.59	184.32	184.21	169.43	169.64
12/6/2017	81.43	82.78	183.88	185.3	169.62	169.01
12/7/2017	82.52	82.49	184.84	185.2	168.76	169.32
12/8/2017	82.57	84.16	185.02	188.54	169.17	169.37
12/9/2017	84.13	84.16	188.44	188.54	169.27	169.37
12/10/2017	83.98	84.16	188.57	188.54	169.3	169.37
12/11/2017	84.05	85.23	188.69	186.22	169.3	172.67
12/12/2017	85.04	85.58	185.83	185.73	172.61	171.7
12/13/2017	85.37	85.35	185.4	187.86	171.43	172.27
12/14/2017	85.27	84.69	187.24	189.56	172.29	172.22
12/15/2017	84.61	86.85	189.46	190.12	172.03	173.97
12/16/2017	86.54	86.85	190.08	190.12	173.72	173.97
12/17/2017	86.68	86.85	189.99	190.12	173.87	173.97
12/18/2017	86.81	86.38	189.69	190.42	173.96	176.42
12/19/2017	86.31	85.83	189.86	187.02	176.36	174.54
12/20/2017	85.61	85.52	186.47	188.82	174.3	174.35

Table 14. Predicted stock close price of tickers compared to their actual stock close price (continued).

Date	GOOG		MCD		FB	
	Predicted	Actual	Predicted	Actual	Predicted	Actual
12/5/2017	999.73	1005.15	170.29	172.99	171.58	172.83
12/6/2017	1006.16	1018.38	172.85	173.48	172.81	176.06
12/7/2017	1018.43	1030.93	173.23	172.91	175.66	180.14
12/8/2017	1031.27	1037.05	172.95	173.15	179.83	179
12/9/2017	1038.28	1037.05	172.88	173.15	179.33	179
12/10/2017	1037.04	1037.05	172.78	173.15	179.04	179
12/11/2017	1035.7	1041.1	172.82	173.25	178.46	179.04
12/12/2017	1039.54	1040.48	173.22	172.23	178.56	176.96
12/13/2017	1038.75	1040.61	171.92	173.55	176.83	178.3
12/14/2017	1039.55	1049.15	173.58	173.14	178.06	178.39
12/15/2017	1048.79	1064.19	172.84	174.06	178.16	180.18
12/16/2017	1063.62	1064.19	173.99	174.06	180.05	180.18
12/17/2017	1063.88	1064.19	173.76	174.06	180.17	180.18
12/18/2017	1064.58	1077.14	173.78	174.2	179.96	180.82
12/19/2017	1076.31	1070.68	174.07	173.39	180.63	179.51
12/20/2017	1069.35	1064.95	173.14	172.17	179.41	177.89
Date	AMZN		BABA		C	
	Predicted	Actual	Predicted	Actual	Predicted	Actual
12/5/2017	1133.99	1141.57	169.32	168.96	77.02	76.54
12/6/2017	1140.02	1152.35	168.67	172.63	76.35	75.44
12/7/2017	1149.25	1159.79	172.29	174.47	75.36	74.98
12/8/2017	1159.53	1162	174.04	177.62	74.85	75.71
12/9/2017	1162.13	1162	177.49	177.62	75.65	75.71
12/10/2017	1160.82	1162	177.39	177.62	75.55	75.71
12/11/2017	1159.29	1168.92	177.06	179.29	75.68	75.85
12/12/2017	1166.17	1165.08	179.1	174.64	75.78	76.15
12/13/2017	1162.9	1164.13	174.26	176.47	76.08	75.14
12/14/2017	1163.65	1174.26	176	171.75	74.98	73.92

Table 14. Predicted stock close price of tickers compared to their actual stock close price.

(continued)

Date	AMZN		BABA		C	
	Predicted	Actual	Predicted	Actual	Predicted	Actual
12/15/2017	1171.43	1179.14	171.44	173.55	73.82	74.77
12/16/2017	1177.69	1179.14	173.25	173.55	74.73	74.77
12/17/2017	1178.99	1179.14	173.45	173.55	74.64	74.77
12/18/2017	1178	1190.58	172.94	173.37	74.72	75.67
12/19/2017	1187.73	1187.38	173.31	171.28	75.63	74.7
12/20/2017	1184.82	1177.62	170.93	172.64	74.59	74.66

Date	GILD	
	Predicted	Actual
12/5/2017	72.99	73.29
12/6/2017	73.22	73.29
12/7/2017	73.22	72.72
12/8/2017	72.73	74.22
12/9/2017	74.46	74.22
12/10/2017	74.33	74.22
12/11/2017	74.33	75.88
12/12/2017	75.99	76.09
12/13/2017	76.07	76.58
12/14/2017	76.76	74.34
12/15/2017	74.35	75.57
12/16/2017	75.52	75.57
12/17/2017	75.48	75.57
12/18/2017	75.6	75.13
12/19/2017	75.35	74.35
12/20/2017	74.38	74.01

4.7. Evaluating the Model

The root-mean-square error (RMSE) is a frequently used measure of the differences between values (sample and population values) predicted by a model or an estimator and the values observed. We evaluate the efficiency of our model predictions looking at RMSE values. RMSE is a measure of accuracy, to compare forecasting errors of different models for a data and not between datasets, as it is scale-dependent. RMSE is the square root of the average of squared errors. The effect of each error on RMSE is proportional to the size of the squared error; thus, larger errors have a disproportionately large effect on RMSE. Consequently, RMSE is sensitive to outliers. We evaluate the accuracy of the models built for every ticker and compare it with the RMSE values of their predictions for 5 days and 15 days. Table 15 has the RMSE comparison matrix.

Table 15. RMSE comparison with predictions between 5 days and 15 days.

RMSE Values of Ticker Model		RMSE Values for Period 12/05/2017 to 12/10/2017		RMSE Values for Period 12/05/2017 to 12/20/2017	
AAPL	1.51	AAPL	1.29	AAPL	0.36
AMZN	7.26	AMZN	7.10	AMZN	7.39
FB	0.98	FB	1.70	FB	2.34
NFLX	2.06	NFLX	1.70	NFLX	1.55
GOOG	8.84	GOOG	7.67	GOOG	7.85
BABA	1.34	BABA	2.41	BABA	2.36
MSFT	0.64	MSFT	0.87	MSFT	0.87
GILD	1.44	GILD	0.93	GILD	0.66
C	0.57	C	0.65	C	0.57
MCD	0.85	MCD	0.9752	MCD	1.15

We observe that RMSE values for both 5 days and 15 days period are in line with the test dataset run on the ticker model, which explain the model are neither overfit or underfit. The RMSE values are higher for AMZN and GOOG and the reason is both these tickers rally with big variation over the period increasing the error rate.

CHAPTER 5. SUMMARY AND CONCLUSION

In this dissertation, we have shown the importance of social media such as Twitter to illustrate the effect of real time news and events on the finance market. In the first portion of this dissertation, we have included a programming method to procure the tweets from investors in real time and store the data in a centralized cloud-based data store. We also discuss in detail the limitations of the Twitter platform and how a developer can code around those limitations. In this research, we have focused on 3000 investors who have been manually handpicked. All the tweets from these investors are pulled in real time and are is stored daily in the centralized data store since 2015/01/01. In this research we focus on the top 10 tickers across various sectors which are mostly been discussed amongst our chosen inventors and we focus our experimental analysis on these 10 tickers.

The second portion of this dissertation focuses on the importance of sentiments in social media and the various parameters to consider while building a sentiment analysis tool. The dissertation focuses on building Microsoft's Azure Sentiment Analyzer and its comparisons with other commercially available sentiment analyzers such as Stanford's NLP. We do various statistical tests and show that the Azure Sentiment Analyzer outperforms the other sentiment analyzers on different datasets.

The third portion of this dissertation focuses on studying the effect of the sentiment scores on predicting the close price of the stock. Once we confirm from experimental analysis, that's sentiments influence the closing price of the stocks we train the Bayesian Linear Regression model for every ticker with the historical stock and sentiment data. In this research we also conclude that the sentiments to influence the closing price of the stock we should consider last 7 days from the

Granger causality analysis. We consider three distinct experiments to choose the time-period which has the best predictive results and conclude from our research that last 7 days including weekends with sentiments considered during the weekends yield the best predictive response. We use the trained Bayesian Linear Regression model for each of these tickers to predict the close price of the stocks for the next 2 weeks and observe that the RMSE values of the predicted values compared to the actual values are less for stable stocks compared to Amazon.com. (AMZN) and Google (GOOG). Selecting the right set of investors and processing the sentiment of those investors through a proper sentiment analysis tools and choosing the right regression model it is possible to predict the closing price of stocks within negligible margin of error.

REFERENCES

[1] Jiawei, H., & Kamber, M. (2001). Data mining: concepts and techniques. San Francisco, CA, itd: Morgan Kaufmann, 5.

[2] Adriaans, P., & Zantinge, D. (1996). Data Mining. Harlow. England: Addison Wesley.

[3] Mostafa, MM. (2013). More than words: social networks text mining for consumer brand sentiments. Expert Syst Appl. 2013;40(10):4241–51.

[4] Yang, C., Fayyad, U., & Bradley, P. S. (2001, August). Efficient discovery of error-tolerant frequent itemsets in high dimensions. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 194-203). ACM.

[5] Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information processing & management, 24(5), 513-523.

[6] Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. Communications of the ACM, 18(11), 613-620.

[7] Baker, M. and J. Wurgler (2006), "Investor Sentiment and the Cross-section of Stock Returns," Journal of Finance,(61)(4), pp. 1645-80.

[8] Baker, M. and J. Wurgler (2007), "Investor Sentiment in the Stock Market," Journal of Economic Perspectives,(21)(2), pp. 129-151.

[9] Bollen, J., Gonçalves, B., Ruan, G., & Mao, H. (2011). Happiness is assortative in online social networks. Artificial life, 17(3), 237-251.

[10] Mao, H., Counts, S., & Bollen, J. (2011, November). Computational economic and finance gauges: Polls, search, and twitter. In Meeting of the National Bureau of Economic Research-Behavioral Finance Meeting, Stanford, CT (Vol. 11, No. 5, p. 2011).

[11] Bollen, J., & Mao, H. (2011). Twitter mood as a stock market predictor. Computer, 44(10), 0091-94.

[12] Bollen, J., Mao, H., & Pepe, A. (2011). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. ICWSM, 11, 450-453.

[13] Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of computational science, 2(1), 1-8.

[14] Mittal, A., & Goel, A. (2012). Stock prediction using twitter sentiment analysis. Standford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf), 15.

[15] Prasad, A. M., Iverson, L. R., & Liaw, A. (2006). Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems, 9(2), 181-199.

[16] Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.

[17] Bishop, C. M. (2006). Pattern recognition. Machine Learning, 128, 1-58.

[18] Raftery, A. E., Madigan, D., & Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. Journal of the American Statistical Association, 92(437), 179-191.

[19] Hamilton, J. D. (1994). Time series analysis (Vol. 2). Princeton: Princeton university press.

[20] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2000). Introduction to the Logistic Regression Model. Applied Logistic Regression.

[21] Friedman, J. H. (2002). Stochastic gradient boosting. Computational Statistics & Data Analysis, 38(4), 367-378.

[22] Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). The annals of statistics, 28(2), 337-407.

[23] Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. Journal of Animal Ecology, 77(4), 802-813.

[24] Shuai, X., Pepe, A., & Bollen, J. (2012). How the scientific community reacts to newly submitted preprints: Article downloads, twitter mentions, and citations. PloS one, 7(11), e47523.

[25] Li, D., Ding, Y., Shuai, X., Bollen, J., Tang, J., Chen, S., ... & Rocha, G. (2012). Adding community and dynamic to topic models. Journal of Informetrics, 6(2), 237-253.

[26] Shuai, X., Liu, X., & Bollen, J. (2012, April). Improving news ranking by community tweets. In Proceedings of the 21st International Conference on World Wide Web (pp. 1227-1232). ACM.

[27] Mao, H., Pepe, A., & Bollen, J. (2010, July). Structure and evolution of mood contagion in the Twitter social network. In Proceedings of the International Sunbelt Social Network Conference XXX, Riva del Garda.

[28] Mao, H., Counts, S., & Bollen, J. (2015). Quantifying the effects of online bullishness on international financial markets. ECB Statistics Paper Series, 9.

[29] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2(1–2), 1-135.

[30] Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Modern information retrieval (Vol. 463). New York: ACM press.

[31] Sayce, David. (2016). Number of Tweets per day. Retrieved from http://www.dsayce.com/social-media/tweets-day/

[32] Granger, C. W. (1988). Some recent development in a concept of causality. Journal of econometrics, 39(1-2), 199-211.

[33] Granger, C. W. (1988). Causality, cointegration, and control. Journal of Economic Dynamics and Control, 12(2-3), 551-559.

[34] Granger, C. W., Huangb, B. N., & Yang, C. W. (2000). A bivariate causality between stock prices and exchange rates: evidence from recent Asianflu☆. The Quarterly Review of Economics and Finance, 40(3), 337-354.

[35] Asur, S., & Huberman, B. A. (2010, August). Predicting the future with social media. In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on (Vol. 1, pp. 492-499). IEEE.

[36] Ariyo, A. A., Adewumi, A. O., & Ayo, C. K. (2014, March). Stock price prediction using the ARIMA model. In Computer Modelling and Simulation (UKSim), 2014 UKSim-AMSS 16th International Conference on (pp. 106-112). IEEE.

[37] Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. International journal of forecasting, 22(4), 679-688.

[38] Arowolo, W. B. (2013). Predicting Stock Prices Returns Using Garch Model. The International Journal of Engineering and Science, 2(5), 32-37.

[39] Das, S., Poggio, T., & Lo, A. Emergent Properties of Price Processes in Artificial Markets. Ret, 10, 3.

[40] Hutchinson, J. M., Lo, A. W., & Poggio, T. (1994). A nonparametric approach to pricing and hedging derivative securities via learning networks. The Journal of Finance, 49(3), 851-889.

[41] Dahan, E., Kim, A. J., Lo, A. W., Poggio, T., & Chan, N. (2011). Securities trading of concepts (STOC). Journal of Marketing Research, 48(3), 497-517.

[42] Chan, N. T., Dahan, E., Lo, A. W., & Poggio, T. (2001). Experimental markets for product concepts.

[43] Xu, S. Y. (2014). Stock Price Forecasting Using Information from Yahoo Finance and Google Trend. URL https://www. econ. berkeley. edu/sites/default/files/Selene% 20Yue% 20Xu. pdf.

[44] Kim, A. J., Shelton, C. R., & Poggio, T. (2002). Modeling Stock Order Flows and Learning Market-Making from Data.

[45] Lo, A., Chan, N., Lebaron, B., & Poggio, T. (1999). Information dissemination and aggregation in asset markets with simple intelligent traders (No. 653). Society for Computational Economics.

[46] Lo, A., Chan, T., & Poggio, T. (2009). U.S. Patent No. 7,599,876. Washington, DC: U.S. Patent and Trademark Office.

[47] Azure Machine Learning Group. (2016). Machine Learning. Retrieved from https://azure.microsoft.com/en-us/services/machine-learning/

[48] Parimi, Nagender. (2015). Introducing Text Analytics in the Azure ML Marketplace. Retrieved from http://blogs.technet.com/b/machinelearning/archive/2015/04/08/introducing-text-analytics-in-the-azure-ml-marketplace.aspx

[49] Dreman, D., S. Johnson, D. Macgregor, and P. Slovic (2001), "A Report on the March 2001 Investor Sentiment Survey," Journal of Psychology and Financial Markets, (2)(3), pp. 126-34.

[50] Thorp, W. A. (2004), "Investor Sentiment as a Contrarian Indicator," The American Association of Individual Investors, Sept.-Oct. 2004.

[51] Shiller, R.J. (2003), “From Efficient Markets Theory to Behavioral Finance,” Journal of Economic Perspectives, (17)(1), pp. 83-104.

[52] Hall, R. E. (2001), "Struggling to Understand the Stock Market," American Economic Review,(91)(2), pp. 1-11.

[53] Malkiel, B. G. and E. F. Fama (1970), "Efficient Capital Markets: A Review of Theory and Empirical Work," The Journal of Finance,(25)(2), pp. 383-417.

[54] Malkiel, B. G. (2003), "The Efficient Market Hypothesis and Its Critics," Journal of Economic Perspectives, (17)(1), pp. 59-82.

[55] Barberis, N., A. Shleifer, and R. Vishny (1998), “A Model of Investor Sentiment,” Journal of Financial Economics, (49), pp. 307-343

[56] Lemmon, M. and E. Portniaguina (2006), "Consumer Confidence and Asset Prices: Some Empirical Evidence," Review of Financial Studies,(19)(4), pp. 1499-529.

[57] Zheng, Y. (2015), "The Linkage between Aggregate Investor Sentiment and Metal Futures Returns: A Nonlinear Approach," The Quarterly Review of Economics and Finance, (58), pp. 128-42.

[58] Kaplanski, G., H. Levy, C. Veld, and Y. Veld-Merkoulova (2014), "Do Happy People Make Optimistic Investors?," Journal of Financial and Quantitative Analysis,(50)(1-2), pp. 145-68.

[59] Ling, D. C., A. Naranjo, and B. Scheick (2013), "Investor Sentiment, Limits to Arbitrage and Private Market Returns," Real Estate Economics,(42)(3), pp. 531-77.

[60] Babu, A. S. and R. R. Kumar (2015), "The Impact of Sentiments on Stock Market: A Fuzzy Logic Approach," The IUP Journal of Applied Finance, (21)(2), pp. 22-33.

[61] Chatterjee, A., & Perrizo, W. (2015, August). Classifying stocks using P-Trees and investor sentiment. In Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on (pp. 1362-1367). IEEE.

[62] Shefrin, H and M. Statman (1985), "The Disposition to Sell Winners Too Early and Ride Losers Too Long: Theory and Evidence," The Journal of Finance,(40)(3), pp. 777-790.

[63] Shefrin, H. (1999), Beyond Greed and Fear: Understanding Behavioral Finance and the Psychology of Investing. Boston, MA: Harvard Business School Press, Revised version published 2002, New York: Oxford University Press

[64] Barber, B.M. and T. Odean (1999), "The Courage of Misguided Convictions," Financial Analysts Journal, (55)(6), pp. 41-55.

[65] Chuang, W.I. and B. S. Lee (2006), "An Empirical Evaluation of the Overconfidence Hypothesis," Journal of Banking & Finance,(30)(9), pp. 2489-515.

[66] Chatterjee, A., & Perrizo, W. (2016, August). Investor classification and sentiment analysis. In Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on (pp. 1177-1180). IEEE.

[67] Tang, H., Tan, S., & Cheng, X. (2009). A survey on sentiment detection of reviews. Expert Systems with Applications, 36(7), 10760-10773.

[68] Cai, K., Spangler, S., Chen, Y., & Zhang, L. (2010). Leveraging sentiment analysis for topic detection. Web Intelligence and Agent Systems: An International Journal, 8(3), 291-302.

[69] Qiu, G., He, X., Zhang, F., Shi, Y., Bu, J., & Chen, C. (2010). DASA: dissatisfaction-oriented advertising based on sentiment analysis. Expert Systems with Applications, 37(9), 6182-6191.

[70] Kim, S. M., & Hovy, E. (2004, August). Determining the sentiment of opinions. In Proceedings of the 20th international conference on Computational Linguistics (p. 1367). Association for Computational Linguistics.

[71] Massa, M. and V. Yadav (2015), "Investor Sentiment and Mutual Fund Strategies," Journal of Financial and Quantitative Analysis,(50)(04), pp. 699-727.

[72] Forsythe, R., Nelson, F., Neumann, G. R., & Wright, J. (1992). Anatomy of an experimental political stock market. The American Economic Review, 1142-1161.

[73] Manski, C. F. (2004). Measuring expectations. Econometrica, 72(5), 1329-1376.

[74] Wolfers, J., & Zitzewitz, E. (2006). Prediction markets in theory and practice (No. w12083). national bureau of economic research.

[75] Berg, J., Forsythe, R., Nelson, F., & Rietz, T. (2008). Results from a dozen years of election futures markets research. Handbook of experimental economics results, 1, 742-751.

[76] Tetlock, P. (2004). How efficient are information markets? Evidence from an online exchange. Social Science Research Network.

[77] Ortner, G. (1998). Forecasting markets—An industrial application. mimeo.

[78] Chen, A. S., Leung, M. T., & Daouk, H. (2003). Application of neural networks to an emerging financial market: forecasting and trading the Taiwan Stock Index. Computers & Operations Research, 30(6), 901-923.

[79] Nagar, A., & Hahsler, M. (2012). Using text and data mining techniques to extract stock market sentiment from live news streams. In International Conference on Computer Technology and Science (ICCTS 2012), IACSIT Press, Singapore.

[80] Yu, C. H., Jannasch-Pennell, A., & DiGangi, S. (2011). Compatibility between text mining and qualitative research in the perspectives of grounded theory, content analysis, and reliability. The Qualitative Report, 16(3), 730.

[81] Shynkevich, Y., McGinnity, T. M., Coleman, S., & Belatreche, A. (2015, July). Stock price prediction based on stock-specific and sub-industry-specific news articles. In Neural Networks (IJCNN), 2015 International Joint Conference on (pp. 1-8). IEEE.

[82] Wolfers, J., & Zitzewitz, E. (2004). Prediction markets. The Journal of Economic Perspectives, 18(2), 107-126.

[83] Breen, J. (2011). R by example: Mining Twitter for consumer attitudes towards airlines. Boston Predictive Analytics Meetup Presentation.

[84] Waugh, Rob. (2012). The Tweets are paved with gold: Twitter 'predicts' stock prices more accurately than any investment tactic, say scientists. Retrieved from http://www.dailymail.co.uk/sciencetech/article-2120416/Twitter-predicts-stock-prices-accurately-investment-tactic-say-scientists.html

[85] Zarrella, Dan. (2009). Is 22 Tweets-Per-Day the Optimum? Retrieved from https://blog.hubspot.com/blog/tabid/6307/bid/4594/Is-22-Tweets-Per-Day-the-Optimum.aspx#sm.00000wkb1cny9bf5utvj1ext1hkbi

[86] TripAdvisor Annotated Dataset. Retrieved from http://nemis.isti.cnr.it/~marcheggiani/datasets/

[87] Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1(12).