Relationship Discovery of Price Movements between Sentiment Analysis on Social Media Data and Stock Market

by

Mohammed Moosa Naqvi

A thesis submitted to
The Faculty of Graduate Studies of
The University of Manitoba
in partial fulfillment of the requirements
of the degree of

Master of Science

Department of Computer Science
The University of Manitoba
Winnipeg, Manitoba, Canada
July 2019

Thesis advisor

Dr. Ruppa K. Thulasiram

Co-Advisor Dr. Milton Boyd

Author

Mohammed Moosa Naqvi

Relationship Discovery of Price Movements between Sentiment Analysis on Social Media Data and Stock Market

Abstract

A desire to make a profit on investment has been a prominent motivational factor in financial investments. The idea of growing with a blue chip firm or an emerging start-up has allured both individual investor(s) and large investing firms alike. One of the financial market areas that gives such opportunity to become part of something bigger is the stock market. Across the globe, stock exchanges become the medium through which billions of stocks are traded on daily basis. Nevertheless, stock market volatility always challenges a seasoned investor to find new ways to invest into stocks that will be profitable in near future. These challenges are equally important for financial firms that are building algorithms for creating profitable stock portfolio. With the advent of social media and similar resonance in digital news media, we have witnessed huge data explosion and this has also opened new opportunities to harvest these data into information for profitable stock trading.

In this research, I have performed analysis of more than 8.5 million news article and twitter messages to determine relationship between stock price and media sentiments. Using novel data visualization and Natural Language Processing techniques,

I have implemented novel data visualizations such as frequency of news items and other related events affecting the company share price.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Stock trading produces a tremendous opportunity to make profit in both long term and short term periods. However, this opportunity comes at the cost of high risk. Reducing risk while making high profit has always been a challenge for investors. Several approaches have been suggested to address this challenge. Sentiment analysis of social media data for stock prediction has emerged as an additional and novel method to address this challenge. Otherwise, most financial firms rely on the success of overall stock market growth for profitable portfolio but these days social media sentiments are having an impact on stock market performance as well.

1.1 Social Media sphere

In the world of social media, each social networking site is serving to a unique user demand or necessity. For instance, main idea behind *Twitter* is to provide a platform for people to provide short and quick update on their life in a limit of 280

characters [8]. This character limit presents an interesting challenge in the field of sentiment analysis, while at the same time as character length is limited, in most cases it is easier to detect user sentiment from message, as message contains precise sentiment keywords without complex sentence structure. Another website, *Reddit* - a self-claimed front page of internet [9] - work on the principle of content quality. This site is divided into sub-reddits and contents are arranged by the number of up votes, with posts with highest up vote on the top. This site is also one of the crucial platform to gauge the trending news. *StockTwits*, a social media website is created on the same concept as Twitter but to specifically cater to the demand on stock market audience. One of the notable feature of *StockTwits* is *cashtag* feature [10]. This feature works on a concept similar to *hashtag* but on this site [10] it has been used for mentioning stock ticker in message. With a two billion user base, Facebook is still dominating the social media universe [11]. This site has been created with the intention of connecting user with his real life family and friends but this site has seen rapid shifts in purpose of use since its inception. Nowadays, it has been used heavily to create impact on user opinion. In 2016 US Presidential election, Facebook had a huge impact in influencing peoples' opinion and debate about the service responsibilities still continues in news media [12].

1.2 Celebrity's influence on social media

In past years, celebrities ranging from politicians to business executive have influenced the stock market in both positive and negative ways. For instance, United States President Donald Trump's tweet expressing disappointment regarding automobile

manufacturing company Toyota's proposal of moving its manufacturing operation from US to Mexico sent its stock value plummeting. This tweet's reaction caused a huge loss of $1.2B [13] in stock value for Toyota. Usually, this kind of influential power is dependent on celebrity fan following in social media. In another example, Kylie Jenner, one of most influential Twitter celebrity, sent SnapChat stock temporarily nosediving 7% after expressing her sadness regarding the inactivity of aforementioned social media application [14]. On the other hand, there are famous social media celebrities whose impact caused positive effects on stock prices as well. John McAfee's tweets about various cryptocurrencies such as SAFEX & Electroneum made their stocks soaring up to 57% [15]. These evidences show the crucial impact of social media on stock market.

1.3 Herd behavior in social media

Impact of social media on public opinion is huge and this impact is not only limited to few social media celebrities with fan following in millions. In various scenarios, it has been observed that herd behavior of common users in social media started a trend and their impact had impacted the companies millions of dollars both in positive as well as in negative ways. For example, Joanne McCoy tweeted about how badly she wanted the Audi R8 V10 with the hashtag #WantAnR8 and later, Audi had spiraled this *hashtag* into a mega social media campaign [16]. Based on these examples it can be realized that there exists both kind of use cases where an individual or certain group of less known people created impact on social media by generating social media trend as well as an individual with huge fan following affecting

the future of companies and consumer products on their own.

1.4 News impact on stock market

Since social media impact on stock market is huge, we cannot ignore the effects that news create on stock prices. As news are created after verifying facts and therefore more credible, its effect on market is more concrete. There are numerous ways that can be used to gain new information. One of the most feasible ways for performing sentiment analysis on news article is by performing crawling on news aggregators sites, such as, RealClearMarkets *(https://www.realclearmarkets.com/)*, StockSnips *(https://www.stocksnips.com/)*. Websites like Bloomberg *(https://www.bloomberg.com/)*, and WallStreetJournal *(https://www.wsj.com/)*, etc., provide news information that shapes markets. Technically, web site crawling makes it possible to read news items without manual assistance. Later, this output can be passed on to sentiment analysis engine for extracting sentiment out of news article.

1.5 Correlation studies

In recent years, there has been immense interest generated for finding correlation between social media sentiments and stock prices, at both professional and academic level. As it is evident from above examples that professional interest exist at several levels, and in this section I emphasize on academic level interests. Nguyen et al. [17] have shown in their research study that the correlation between 18 major US

stock prices such as Apple, Amazon, Ebay, Dell, etc., and message sentiment on each company's respective message boards. Their proposed model outperformed other methods in the average accuracy of 18 stocks. At Stanford, Mittal and Goel [18] has proved correlation between twitter message sentiments and DJIA (Dow Jones Industrial Average) stocks. The recent study by Chen et al. [19], has demonstrated that the topic-based public mood from Weibo can be used for sentiment anaylysis in Chinese stock market.

In terms of news media sentiment analysis and its correlation with stock prices, the past research has shown mixed results. In some research studies [20], market has shown over-reaction to news article sentiments while there also exists research studies, that shows market under-reaction to recent news. Moreover, few research studies exist that show market is completely unaffected by news information [21]. All three research studies have proved different hypotheses even without contradicting each other with respect to news content. The core difference between these studies can be found in the field of behavioural finance. As researchers suggested that sometimes investors seems to be self assured of the newly acquired information by confidential sources and later, when same information broke out in public as news article they tend to overreact on same information [20]. The under-reaction theory based on the hypothesis that investors have their individual opinions and interaction between them lead to asymmetric information that results in the state of under reaction [22]. With regard to third theory, established the fact that stock price deviation are mainly caused by mis-specification of asset pricing models [21].

1.6 Evolution of sentiment analysis for different purposes

Although social media catalyzed the trend of sentiment analysis but academic research on core principal of public opinion mining can be dated back to pre world-war era where it has been used with political motivation [23]. In recent years, we have seen a massive increase in the number of papers focusing on sentiment analysis and opinion mining. An early example of academic research based on computer based system can be seen in mid 90s, when Sandri et al. [24] used a computer system for expert opinion analysis in the domain of industrial safety that allowed, for example, a pooling of opinions. According to Mantayla et al. [25], nearly 7,000 papers of sentiment analysis topic have been published and, more interestingly, 99% of the papers have appeared after 2004, making sentiment analysis one of the fastest growing new research areas.

1.7 Herd behaviour in financial market

The field of behavioural finance emphasize heavily on financial market's herd mentality. Both 1990's dot-com bubble and 2008 economic crisis had one thing in common: herd behaviour. In both cases, investors and money managers betted heavily on seemingly promising industries and financial instruments without analyzing the underlying financial models and their sustainability. Austrian economist Joseph Schumpete, [26] drew inferences relating to the psychology of investor behaviour during this period based upon their empirical analysis. In his research, he concludes

that appreciation of the psychological inputs to investment decisions could have assisted financial managers in avoiding serious mistakes and enable them to construct more viable investment strategies. Famous investor, Warren Buffet, has aptly warned against herd behaviour mentality in his quote which has become a *mantra* for value investors.

"And if they insist on trying to time their participation in equities, they should try to be fearful when others are greedy and greedy only when others are fearful." [27]

As it is evident from above discussion that fields of finance, social media and sentiment analysis has vast research universe in itself. There is a plethora of research opportunities that still need to be explored by combining or utilizing the past research as a bed rock for the ground of new discoveries.

The following sections describe the related work (Section 2) and discusses the problem statement (Section 3). Later, the experimental framework is explained in detailed manner (Section 4) with the evaluation of the proposed method (Section 5).

Chapter 2

Related Work

Stock market sentiment analysis has been among one of the most lucrative domain, which if mastered could be highly profitable. Usually, this kind of sentiment analysis has been performed by a team of highly experienced experts with both financial and technical knowledge but in recent years algorithmic trading is entering into this domain as well. Use of advance data science and machine learning techniques in recent past years has shown promising results.

2.1 Machine Learning and Artificial Intelligence Based Sentiment Analysis Research

Today, sentiment analysis looks relatively simple and works extremely well; this was achieved after significant efforts by researchers who have invented different methodologies and attempted numerous models. The advances in field of machine learning and artificial intelligence has also contributed immensely in the arena of

sentiment analysis. The new pinnacle of sentiment analysis research today have the advantage of knowledge gained by years of research in respective fields. In this section, I would like to present commonly used machine learning and artificial intelligence based sentiment analysis techniques.

2.1.1 Attention-Based Neural Networks

Lin et al. [1] proposed a new model for extracting an interpretable sentence embedding (set of techniques in NLP where sentences are mapped to vectors of real numbers) by introducing self-attention. Instead of using a vector, they have used a 2-D matrix to represent the sentence embedding, with each row of the matrix attending on a different part of the sentence.

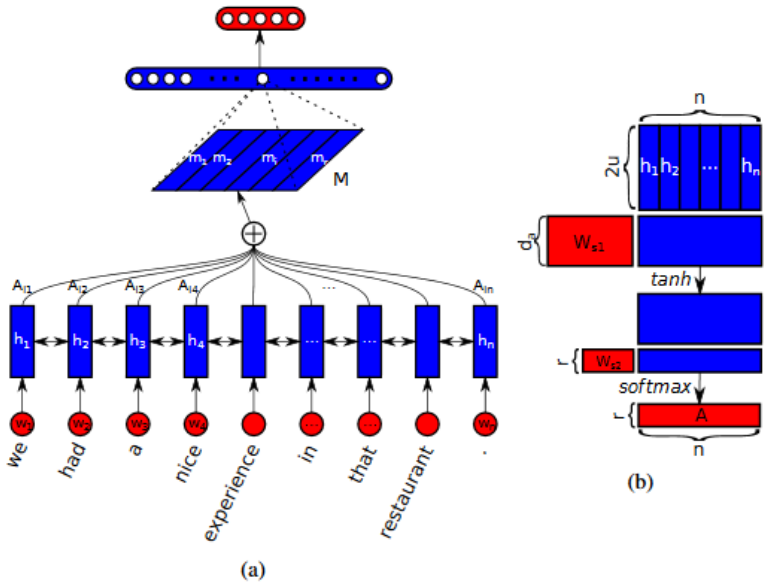

Figure 2.1: [1] A sample model structure showing the sentence embedding model (a). a bidirectional LSTM, where the summation weights are computed in a way illustrated in (b). Self Attention mechanism.

The proposed sentence embedding model consists of two parts. The first part is a bidirectional Long Short-Term Memory (LSTM) model, and the second part is the self-attention mechanism, which provides a set of summation weight vectors for the LSTM hidden states. These set of summation weight vectors are dotted with the LSTM hidden states, and the resulting weighted LSTM hidden states are considered as an embedding for the sentence.

Although researchers proposed attention-based neural network (NN) on top of the convolutional neural network (CNN) or LSTM model to introduce extra source of information to guide the extraction of sentence embedding ([28]), the former can not be fit for analyzing sentiment and related tasks, where the output is one attribute for the entire sentence i.e. positive, negative or neutral.

2.1.2 Recursive Neural Tensor Network

Socher et al. [2] in their research of sentiment analysis have constructed sentiment tree based on recursive deep neural network (DNN). In practice, semantic word spaces have been very useful but cannot express the meaning of longer phrases in a principled way. They have introduced a Stanford Sentiment Treebank, which consists of grained sentiment labels for 215,154 phrases in the parse trees of 11,855 sentences. The authors introduced the Recursive Neural Tensor Network (RNTN), which was trained on Standford Sentiment Treebank. The model outperformed all previous methods on several metrics and pushed the state-of-the-art in single sentence positive/negative classification from 80% up to 85.4%. A more advanced version of this algorithm called Tree LSTM was proposed by Tai el al. [29] in 2015. The Tree

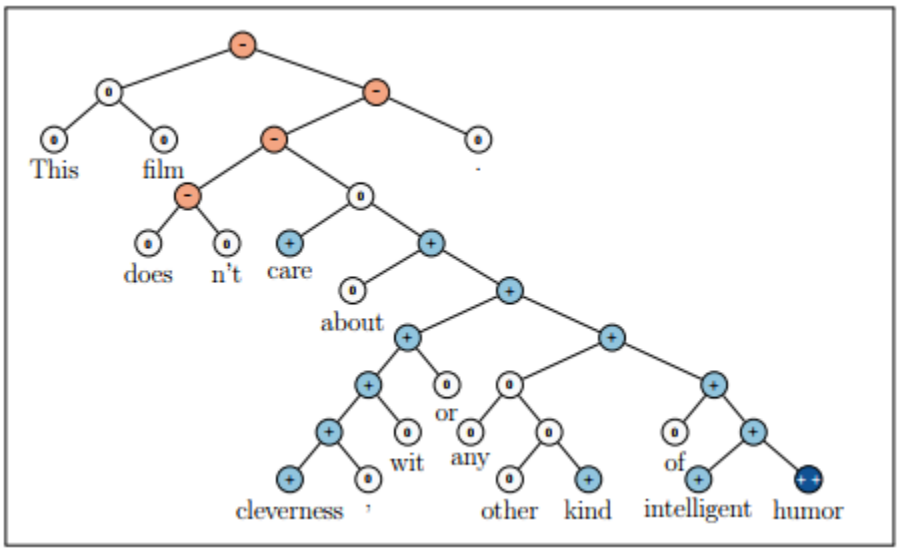

Figure 2.2: [2] Example of the Recursive Neural Tensor Network accurately predicting 5 sentiment classes, very negative to very positive (, , 0, +, + +), at every node of a parse tree.

LSTM model is a generalization of LSTMs to tree-structured network topologies. TreeLSTMs outperform all existing systems and strong LSTM baselines on sentiment classification on Stanford Sentiment Treebank dataset. Figure 2.2 shows the Recursive Neural Tensor Network proposed by Socher et al. [2].

2.1.3 Unsupervised Sentiment Neuron

In 2017, Machine Learning research lab OpenAI, founded by Elon Musk, developed an unsupervised system for sentiment analysis. Radfor et al. [3] explored the properties of byte-level recurrent language models. In their research, they have proved that a tagged sentiment corpus is not required to train a supervised sentiment model. They showcased a normal character level Recurrent Neural Network (RNN) can figure out the positive or negative sentiment on its own. For training purpose, they

have used the same Stanford sentiment Treebank dataset that has been mentioned in previous subsection. It is important to mention that the primary objective of this research is to generate product reviews based on Amazon product reviews given by real users. As shown below, the sentiment analysis of RNN generated review has been successful to a high degree of accuracy.

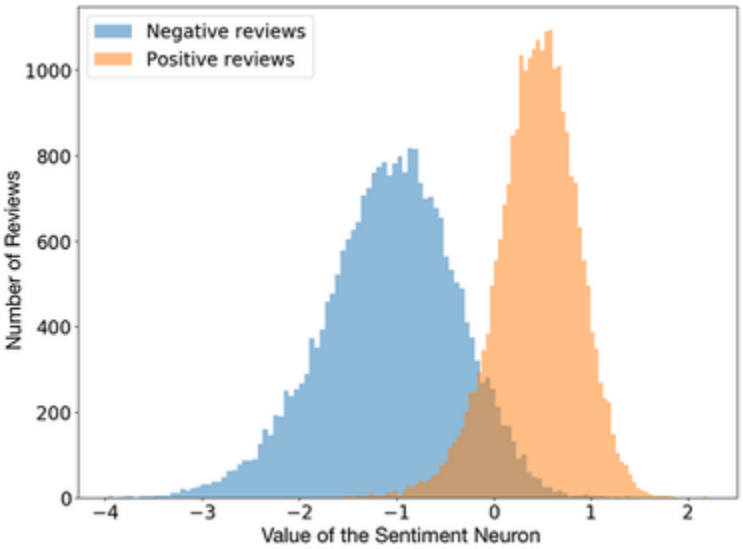

Figure 2.3: [3] The sentiment neuron within model can classify reviews as negative or positive, even though the model is trained only to predict the next character in the text.

In Figure 2.3 unsupervised sentiment neuron model has been able to successfully predict the reviews sentiment polarity for the most part with only few exceptions.

2.2 Social Media Sentiment Analysis for Stock Prediction

A lot of phenomenal work has been done on sentiment analysis for stock price prediction. Qasem et al. [30] have worked on twitter data to predict stock for four major US companies Google, Twitter, Facebook and Tesla. Their work computed

the sentiment of social media messages. These messages were filtered from Twitter based on emoticons.

Houlihan and Creamer [31] have shown that sentiment analysis of social media data combined with market trading data improves stock market prediction. To arrive at their results, they have used various machine learning techniques such as Naïve Bayes, Logistic Regression, etc.

Ceron et al. [32] showed in their research study the existence of good correlation between user sentiments on Twitter and their political preferences. Although internet users are not necessarily representative of the whole population of a country, this analysis shows a remarkable ability for social media to forecast electoral results.

Nguyen et al. [17] research study showed that social media can be used to predict stock market movement. Unlike previous approaches where the overall moods or sentiments are considered, in this approach, the sentiments of the specific topics of the company are incorporated into the stock prediction model.

According to Bloomberg [33], social media and news sentiment have more impact on small-cap stocks as compared to large-cap or mid-cap stocks. This market capacity is defined by revenue generated in market by a company. This approach is based on the same Bloomberg research study. Figure 2.4, taken from Bloomberg [4], presents stock portfolio based on sentiment analysis of news and social media and other strategies that are based solely on news or social media.

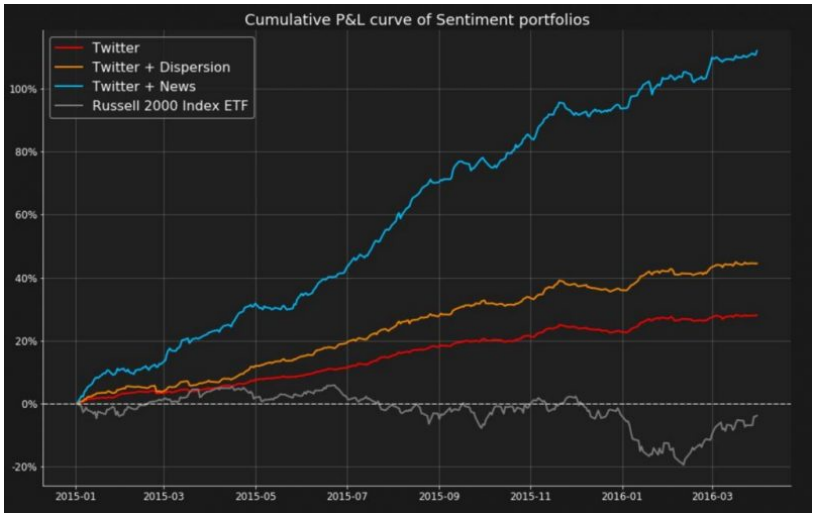

Figure 2.4: [4] Russell 3000 stock portfolio performance based on sentiment analysis of News and Social Media and its comparison with market index

Bollen et al. [5] have shown in their research that collective mood states derived from large-scale Twitter feeds are correlated to the value of the Dow Jones Industrial Average (DJIA) over time. Their finding include an accuracy of 87.6% in predicting the daily up and down changes in the closing values of the DJIA. Although, this paper has been cited widely in academia and news media alike, it consist of some serious flaws. For example, the accuracy has been reported as 87.6% including in the result section, but as result are based on 15 days of test data and prediction was correct on 13 out of 15 days, actual number for accuracy would be 86.7%. Another factor, as 15 days is very short time period, correct prediction on 12 or 14 days would have changed this number by 7 percent.

DJIA daily prediction using SOFNN.

Evaluation	I_{OF}	I_0	I_1	$I_{1,2}$	$I_{1,3}$	$I_{1,4}$	$I_{1,5}$	$I_{1,6}$
MAPE (%)	1.95	1.94	1.83	2.03	2.13	2.05	1.85	**1.79***
Direction (%)	73.3	73.3	**86.7**	60.0	46.7	60.0	73.3	80.0

Figure 2.5: [5] Daily Prediction based on unique sentiments combination

As it is evident from the Figure 2.5, the predicted accuracy holds true in only one case out of 8 predictive models, while other models have shown significantly low accuracy that went down to 40%.

2.3 News media sentiment analysis for stock prediction

Chen and Lu [34] have shown that volume of options trade can be correlated with the stock prices. These research utilized data from CRSP and Option-Matrics implied volatility surface sources.

Gruhl et al. [35] showed that blogs and other on-line social media websites are predecessors to real-world behavior and the volumes of posts related to various products on Amazon website are highly correlated with actual purchase decisions. Pang and Lee [36] provided further support for social media data as a viable source to use in predictive analytics, which is validated by the fact that people are more inclined to share their opinions on social media websites to mere strangers. Extracting features from social media messages have proven to be a robust method for a variety of different labels [36].

Yu et al. [37] study was based on a sample of 18,497 bad news articles

and time series of 1,008 *Russell 3000* stocks returns during the period 2005 to 2017. In their experiment, they employed a time series clustering technique on cumulative abnormal returns of stocks. For input purpose, a dataset of bad news events and stock prices had been used for input. During research, news articles related to those stocks were grouped into different clusters. Later, they applied natural language processing and multi-class classification algorithms such as Support-Vector Machine model (SVM), on relevant news articles to extract features of each cluster. The overall accuracy scores by SVM were not high, most of which were went just slightly over 0.5 in range between 0 to 1.

Chapter 3

Problem Statement

In stock markets highly volatile environment, predicting the stock price always poses a challenge for stock analysts. In the age of digital media, public opinion is believed to have an impact on stock prices. Although, there has been extensive research done in this field, there is still a lot of opportunities to gain insight.

For my thesis, I propose to perform sentiment based analysis to find a relationship between real stocks prices and social media and news articles driven stock prices. Also, I propose to perform an analysis to identify the frequency distribution of companies as well as related topic that may effect the share price of company.

Chapter 4

Solution Methodology

The solution methodology can be divided into following main stages, each stage can be divided into small sub steps, that will be explained in detail in this section:

1. Data Collection
2. Data Gathering
3. Sentiment Analysis
4. Algorithm Implementation
5. Histogram
6. Time Series Analysis
7. Sentiment Words Cloud

4.1 Data Collection

The data sources can be divided into following categories:-

1. **Social media messages**
2. **News headlines**
3. **Stock Prices and Quotes**

4.1.1 Social media messages

For gathering social media messages, I have chosen Twitter as a choice of social media. The Twitter data set that I used for experimentation for my research has been obtained from Google owned site Kaggle. This data set comprises of tweets about stock that are being traded on NYSE and NASDAQ. In particular, this data set deals with 583 publicly traded companies.

Moreover, this data set also focus on the social media personalities influential factor i.e., how much impact is created by tweets of mentioned celebrities. One of the direct ways to gauge this influence is by comparing the social media celebrity's followers count. This data set mostly comprises of Twitter handles who have followings in millions with some handles with followings in thousands. This twitter handles cover wide range of social media presonalities, starting from famous news media outlets such as MarketWatch, YahooFinance, Forbes, WSJ etc, to individual billionaire investors such as Mark Cuban, Ben Bernanke, Jim Cramer etc. This kind of diverse data set helped my research to cover influential social media personality with different set of fan following.

4.1.2 News Media Messages

For the purpose of doing sentiment analysis of news media articles, I have found Reuters news headlines data set to be most vast with high density of news article. This data set contains 8.5 million news headlines spanning around 10 years time period starting from 2007 to 2016. On average, it contains approximately 230 news article for every day through out 10 year time period.

During my research, I have come across other news data sets as well but they were not as vast and dense as above mentioned data set. Apart from obtaining Reuters dataset, I had also contacted Bloomberg from their Bloomberg terminal services at Asper Business School but they have not entertained the request of providing their news dataset. These days this kind of news data sets have become expensive commodity to have, and obtaining them for free of cost has become highly challenging task and buying these sets was not possible due to lack of fund. The Reuters data set used in this research was also unavailable for a period of time before it became available publicly again.

4.1.3 Stock Prices and Quotes

The stock prices were downloaded from MarketWatch website. I have downloaded the stock price based on the stock tweets or news pattern that discovered during the research. For instance, during the launch of an Apple product news media would be exclusively covering the product feature and providing reviews and user response on their site. During data analysis I have found that this brief period of time may range between couple of days up to a week. Based on this discovered time pattern, stock

prices have been picked from MarketWatch site on their respective days.

Another challenge during my research was tagging 8.5 million news articles from Reuters web site. As this record were not tagged, I have downloaded ticker list from BeatTheMarket.com for Russell3000 stocks. Although, Reuters news articles cover more than Russell3000 stocks but expanding research to cover everything was out of scope. I have tried to focus primarily on stocks that are in S&P 500.

4.2 Data Characterstics

During my research thesis, I have worked on three different method of gathering data:

1. Social Media Messages Streaming
2. News Site Scraping
3. Acquiring data sets for social media and news media

4.2.1 Social Media Messages Streaming

For acquiring social media messages, the first choice was streaming messages to get real time data and consequently, near real time sentiment and stock analysis. Usually, social media sites provide their proprietary APIs to extract messages from their platforms. These APIs can only be accessed through application that needs to be created on respective social media site's developer platform. During this phase, I have implemented streaming APIs for two platforms: Twitter and StockTwits.

1. **Twitter**

   Twitter provide its free streaming API with the limitation of 1% of actual generated data. For the API implementation, I was required to create Twitter application to generate Consumer Key, Consumer Secret, Access token and Access Token Secret, values that are necessary to access streaming APIs. Twitter have certain limitations on APIs which made it difficult to use for this research. As mentioned above, rate limiting is one of the biggest challenge in gathering real user sentiment. Besides, Twitter APIs also imposes 15 minute window limit for only 15 calls [38]. This limitations made it difficult to rely on the streaming output.

   However, there are other paid alternatives available in the market, which provides data without any limitations. During my research, I have contacted Twitter for paid APIs. In their response, they mentioned that Entry-level data sets from 1 million Tweets over a 40-day period are priced at $1,250 USD. This pricing is inelastic until either threshold is exceeded. This high price and short time period made this solution quite expensive. There are paid third party data providers, such as GNIP and Datasift, who also provides data from Twitter firehose. Due to lack of funding this option was also not feasible.

   I have implemented Twitter API using Java. During experimentation, I have found that although message amount were limited according to Twitter rate limit even then APIs were giving decent amount of messages for popular stocks. The main issue was there were a lot of noise in collected data resulting in a poor quality of overall data set. Fixing this issue is a highly challenging task

and thus makes it out of scope for this research study.

2. **StockTwits**

   Like Twitter, there is another social media available only to cater to the demand of investment market, it is called StockTwits. This social media site provides its own API similar to Twitter but with more tight limitations. This API can only allow 5 keywords per request, which is very small compared to 400 keywords Twitter API limits.

   I have implemented this API in Java. After implementation, I have found that messages were few in numbers, even for popular stocks such as Apple and Facebook. I have also tried to contact their developer team in regards to information about paid API but I was unable to obtain information from their side. This discovery made it not suitable for further research.

4.2.2 News Media

In the field of finance, news is the big source of insights. However, it is not always possible to read every newspaper and every article manually. Hence, web scraping is used for extracting the valuable inputs from different news stories, headlines etc. to convert them into actionable investment insights.

Initially, I had built website scraping tool in Python using Scrapy framework. In the beginning, I was able to scrap major financial news sites such as Bloomberg, Reuters, MSNBC and Wall Street Journal. Later, due to change in policies of these sites scraping became futile for different reasons. For example, WSJ has made subscription compulsory, that resulted in news article hidden behind pay wall.

This feature made implemented scrappy framework useless. I would like to mention that, scrappy framework can be used to create module that can be used to access content behind pay wall but the cost of subscription and implementation time made it out of scope of this research. Similarly, Bloomberg and Reuters change their XML site pattern that made Scrapy unable to parse those sites. The newer pattern used dynamic ids for HTML elements, which made it harder to parse using Scrapy framework. Due to these new restrictions, I had to abandon the idea of web site scraping.

4.2.3 Acquiring data sets for social media and news media

During my literature review, I have come across papers [30], [5], that have used data sets instead of scraping from websites or using streaming in case of social media. Although, in some cases, the method of obtaining data sets is not mentioned in these papers, it can be assumed that archived data sets have been used as data expands for several years or contains millions of record.

During my research, I have analyzed several data sources and after careful consideration came up with the two data sources that are mentioned in previous section i.e. 27,000 Financial Tweets and 8.5 million Reuters new headlines. These two data sources has been used to cover two different vertical i.e. Social Media and News Media. Based on the Bloomberg study [4] mentioned above, I had decided to choose these data sets. According to the study, it takes considerably more time for market to digest social media sentiment for small cap stocks while news have been proved to create more impact for large cap stocks.

Twitter's financial tweets data set has been obtained from Kaggle [39]. This data set provides two unique advantage over other social media data set that I came across during my research. First, it has manually tagged tweets with stock quotes, which made it easier to classify to specific stock. Second, this data set has been created by taking the social media popularity of person in account.

Reuters data set has been obtained from GitHub. The main motivation behind selecting this data set over other news media data set was densely packed news headlines and coverage of long time period of 10 years. During my research, I was unable to find any other free data set which had spanned for such a long time.

4.3 Sentiment Analysis

During my research, I focused on two different kinds of sentiment analysis i.e. Naive Bayes classification and Natural Language Processing. I started my research using Naive Bayes classification with the main motivation of its easier implementation of algorithm without compromising on result accuracy. Gradually, as my research progressd, based of the result accuracy, I have come to realization that a better alternative must be required. In the quest of better alternative, I have came across Stanford CoreNLP tool. Basically, this tool had covered the shortcomings of Naive Bayes classification of lack of pre-defined training data set. The detailed explanation of both algorithms and their method used in my research is presented below.

4.3.1 Naive Bayes Classification

In my initial attempt, I have used Naive Bayes classification along with streaming data from Twitter, later abandoned due to lack of data. The main idea behind Naive Bayes algorithm is to perform probabilistic classification in two or more categories as per requirement. In my case, I have divided messages into two categories i.e. positive sentiment and negative sentiment. For probabilistic classification, Naive Bayes requires a test data set to generate features set. Later, this feature set has been used to perform classification for real time data. The test data set consisted of 6000 Twitter messages manually classified into positive and negative sentiments. The resulted feature set has been used for message classification. The Naive Bayes algorithm has been implemented in Python framework using *nltk* library. During testing, the Naive Bayes perform with a quite high accuracy that reached up to 70%. Although, this figure considered on the higher side in case of Naive Bayes classification, this accuracy can be improved by performing another kind of sentiment classification.

4.3.2 Sentiment Analysis using Natural Language Processing (NLP)

The primary limitation for Naive Bayes classification is its inability to perform sentiment analysis based on sentence structure. Another issue associated with the algorithm is dependency on similarity between test data set and actual data set. Sometimes, test data set fails to cover features of real time data set. Hence, it results in low accuracy. This dependency issue can be easily addressed by using data sets

that is sufficient to cover all feature of real data set.

In order to overcome above mentioned limitation of Naive Bayes algorithm, my focus shifted on Stanford CoreNLP. It has covered the shortcomings of Naive Bayes algorithm. Socher et al. [2] have introduced a Stanford Sentiment Treebank, which consists of grained sentiment labels for 215,154 phrases in the the form of parse tree. The NLP aspect of algorithm also addresses the sentence structure limitation of former algorithm. In my research, I have used Stanford CoreNLP by importing Java library on Maven architecture.

4.4 Algorithm Implementation

Based on different available data sets, I have built two different algorithms. One for social media messages data set and another for news headlines data set. In the following section, I explain them in detail.

4.4.1 Social Media Message data set

For social media data set, collected data set consists of 28,000 tweets message. These messages are in CSV format. For the sake of clarity I am presenting JSON representation below:

Listing 1 Example Tweet

```
{
"id": "1019719465095790600",
"text": "David's favorite stock isn't Alphabet but Facebook",
"timestamp": "Wed Jul 18 23:04:00 +0000 2018",
"source": "jimcramer",
"symbols": "FB-GOOGL-GOOG",
"company_names": "Facebook*Alphabet*Alphabet",
"url" : "http://bit.ly/2NrYxje",
"verified" : "True"
}
```

The tweets in data set are tagged with actual stock quote that made it more accurate for classifying to specific records. The algorithms developed to find relationship has been explained as follows :

Algorithm 1: Algorithm for social media data

Input : File with lines of tweets meta data in CSV format

Output : None

$totalMessageCount \leftarrow 0$

while *File contains unread CSV line* **do**

```
/* Total message count                                          */
```

$totalMessageCount \leftarrow totalMessageCount + 1$;

Retrieve stock quote list from CSV line;

Split stock quote list with "-" as seprator keyword;

for $prsentStockQuote \leftarrow stockQuoteList$ **do**

$presentStockDate \leftarrow$ Retrieve date string from CSV line;

$presentMessage \leftarrow$ Retrieve text message from CSV line;

$presentSentimentScore \leftarrow$ Retrieve sentiment score from custom built sentiment analysis on top on Stanford CoreNLP;

if *record exists for present stock quote on same date* **then**

```
/* for every message calculated sentiment score value
   will lie in between -1 to 1                                 */
```

$totalSentimentScore \leftarrow totalSentimentScore + presentSentimentScore$

Update message list by adding present message in list;

```
/* In MongoDB database(unique for each stock quote on
   specific date                                               */
```

$totalMessageCount \leftarrow totalMessageCount + 1$;

Update record in MongoDB database;

else

```
/* Create object required in MongoDB format                    */
```

$stockquote \leftarrow presentStockQuote$;

$date \leftarrow presentStockDate$;

$sentimentScore \leftarrow presentSentimentScore$;

$messageList \leftarrow presentMessage$;

Create record in MongoDB;

end

end

end

The above algorithm can be divided into three sections for a better

understanding:

1. **Conditional Section:** As shown in **Listing 1** above, the social media messages are manually tagged with respective company mentioned in tweets. The sole objective of the conditional section is to check for existence of unique record in the database, based on company name and message posted date as key. Based on the result, record is either been inserted or updated after assigning required values accordingly.

2. **Repetition Section:** This ensures that message get appropriately classified to multiple symbols mentioned in JSON record.

3. **Termination Section:** This ensures that every message from the CSV file has been read and message count has been incremented accordingly.

4.4.2 News media data set (Reuters)

For news media data set, collected data set consists of 8.5 million news headlines record. These messages are in CSV format. For the sake of clarity I am presenting JSON representation below :

Listing 2 Example record for news media data

```
{
"timestamp": "10197194650957906 00",
"text": "David's favorite stock isn't Alphabet but Facebook",
"url" : "http://bit.ly/2NrYxje",
}
```

In comparison with Twitter data set, this data set provides less information. Nonetheless, this information is sufficient for my research requirement. The major challenge in this data set was untagged data. To overcome this challenge, I have created a map of S&P 500 stocks and used it for dynamic mapping. The algorithm is mentioned on following page:

Algorithm 2: Algorithm for news media data

Input : File with lines of news meta data in CSV format

A map $StockQuoteMap$, consists of Stock Quotes and Stock Name for S&P 500

Output : None

$totalMessageCount \leftarrow 0$

while *File contains unread CSV line* **do**

`/* Total message count */`

$totalMessageCount \leftarrow totalMessageCount + 1$;

$messageKeywordList \leftarrow$split message by " " as keyword;

$isNewsContainsStockName \leftarrow$ false;

for $keyword \leftarrow messageKeywordList$ **do**

if $keyword$ *presents in* $StockQuoteMap$ **then**

$isNewsContainsStockName \leftarrow$ true;

$presentStockQuote \leftarrow keyfrom$StockQuoteMap $break$

end

end

if $isNewsContainsStockName \leftarrow$ *true* **then**

$presentStockDate \leftarrow$ Retrieve date string from CSV line;

$presentMessage \leftarrow$ Retrieve text message from CSV line;

$presentSentimentScore \leftarrow$ Retrieve sentiment score from custom built sentiment analysis on top on Stanford CoreNLP;

if *record exists for present stock quote on same date* **then**

$totalSentimentScore \leftarrow totalSentimentScore + presentSentimentScore$

Update message list by adding present message in list;

$totalMessageCount \leftarrow totalMessageCount + 1$;

Update record in MongoDB database;

else

$stockquote \leftarrow presentStockQuote$;

$date \leftarrow presentStockDate$;

$sentimentScore \leftarrow presentSentimentScore$;

$messageList \leftarrow presentMessage$;

Create record in MongoDB;

end

end

end

Same as Social Media algorithm, the above algorithm can be divided into

following sections for a better understanding i.e.

1. **Termination Section:** The *while loop* is to ensure that every news article from the CSV file has been read and message count has been incremented accordingly. At this stage, message is also being tokenized for further processing.

2. **Iteration Section:** The tokenzied message from the previous step, passed to *for loop* for classifying it to appropriate company. As this data set is untagged, these step is responsible for message classification. Once message has been identified to a S&P 500 company, algorithm will store company name temporarily for further processing and loop will terminate.

3. **Conditional Section:** The are two *if statement* in this algorithm. The *outer if statement* is responsible for checking if the news article belong to S&P 500 stock, if not, algorithm will proceed to read next message. If news article belong to S&P 500 stock, algorithm calculate sentiment score and read appropriate values from CSV file and send it to *inner if statement*. The sole objective of *inner if statement* is to check for existence of unique record, based on company name and message posted date as key, in database. Based on the result, record is either been inserted or updated after assigning required values accordingly.

4.5 Data Analytics

In the following subsections, the main objective would be to focus on visual representation of data. One of the important aspect of machine learning is about reading patterns in numbers. In these section, the main motivation behind choosing these visualization is as follows:

1. **Histogram**: Analyzing data distribution over extended time period
2. **Time Series**: Finding relationship between stock price and sentiments
3. **Sentiment word clouds**: Showing occurrence of sentiment words with respect to their frequency

In-depth description of these methodologies are as follows:-

4.5.1 Histogram

At the most basic level, a histogram is a graph representation of frequency. It shows how frequently certain values occur in the data. The nice thing about histograms is that they can be used to show the frequency of any type of variable. A sample histogram is shown below generated for representation purpose.

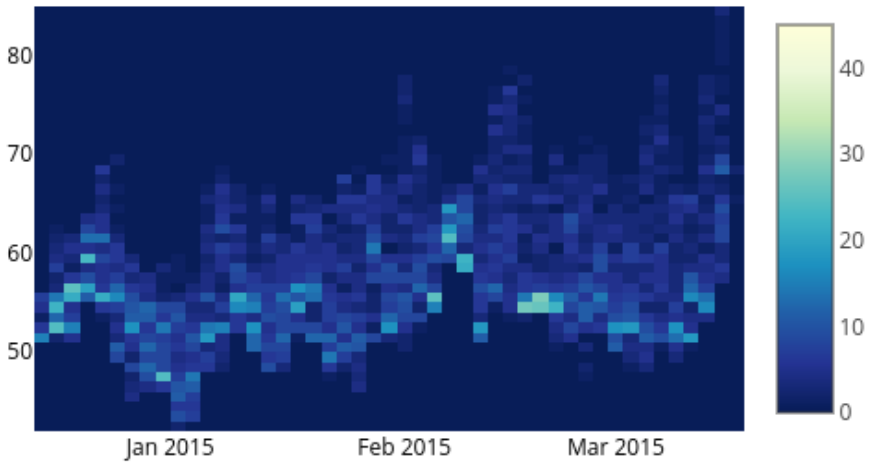

Figure 4.1: [6] Histogram representing temperature changes based on three month time period from January 2015 to March 2015

These temperature values taken at 30 minute time interval has been shown through out the time period of three months.

For my research, I have constructed histogram to display the frequency distribution of news media headline or tweets about stock in observation.

4.5.2 Time Series analysis

Time series is a sequence of data points in a chronological sequence, most often gathered in regular intervals. This regular intervals may typically vary from hourly, daily or weekly depending on the objective. In rare scenarios, such as, monitoring user click patterns, this interval can be at seconds or minute level as well. In general, time series analysis can be applied to any variable that changes over time. Moreover, a time series analysis can be used to find relationship between multiple factors over a period of time. A sample time series analysis representing relationship has been shown below.

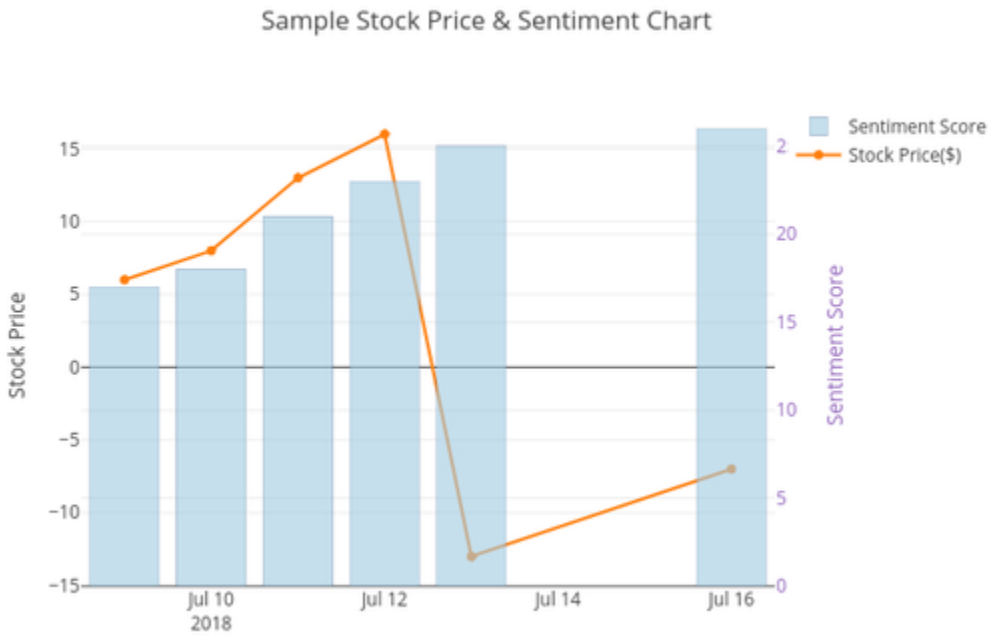

Figure 4.2: Time series representing relationship between stock prices and sentiments over time. Drawn for the illustration purpose using python and plotly library

Normally, the values will not be regular, as sometimes, a stock will not

be mentioned in news or tweets everyday or market will be closed on weekend. Nevertheless, a pattern can be identified regardless of few missing data points.

4.5.3 Sentiment Word Cloud

Word clouds are one of the simple yet crucial aspects of data visualization. The basic concept behind word clouds is to show word's frequency in a cloud form. The size of word present in data cloud is directly proportional to that of word's frequency in the data set. A sentiment word cloud has been used to show word frequency along with the sentiment associated with it. In the Figure 4.3 [7] word clouds showing word frequency in tweets that are associated with Adidas and Nike.

Figure 4.3: [7] Sentiment Word Cloud for Adidas and Nike representing positive and negative sentiment words

In my research, I have performed sentiment analysis on news media headlines for the top five mentioned companies.

Chapter 5

Results and Discussions

During my research, I have analyzed Twitter as well as Reuters data. As Reuters data set was significantly bigger in size, the focus of this result section will be primarily on latter data set. The experiments were performed using Python as a language and Plotly as a visualization tool. The messages were stored in MongoDB, a NoSQL database. The messages were aggregate at daily level uniquely identified by stock name and date. With the main objective to understand the most popular business sectors in news media, I aggregated top 10 companies in their respective business sector and presented in Figure 5.1.

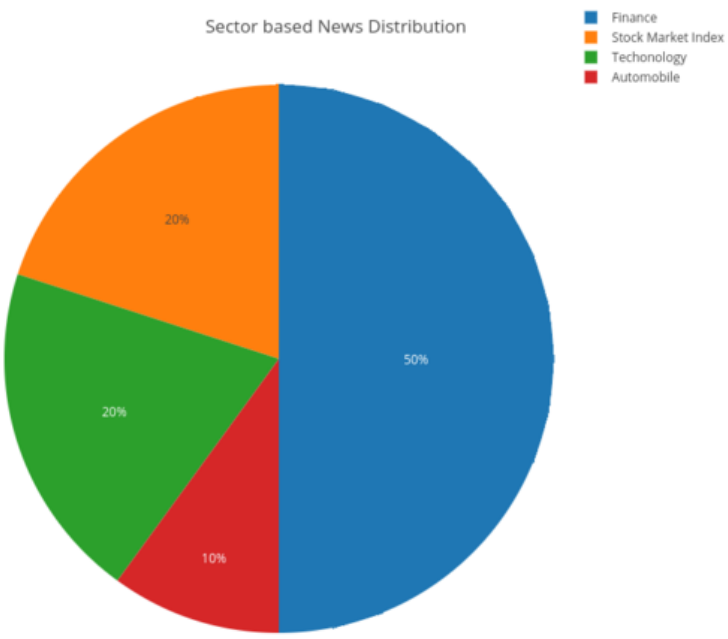

Figure 5.1: Pie Chart representing business sectors based on Top 10 most mentioned companies

The above pie chart has been built on ten years of Reuters news data set from 2007 to 2016. As it has been evident from above figure that financial sector has been dominated the news media. The top five financial sector companies include MSCI, Morgan Stanley, Goldman Sachs, CitiGroup and BlackRock. Both technology sector and stock market index share same number of stocks tickers mentioned with two entities each. In technology sector Microsoft and Apple were mentioned most whereas stock market index was dominated by NASDAQ and Moody's index. In automobile sector, Ford Motors was the only company. The detailed analysis of news stories is explained in the next sub section in form of histograms to provide better understanding of data distribution.

The experimental set up for result has been performed on a single machine

with two different environment i.e., Windows and Linux. Windows environment has been used for development purpose, whereas Linux environment has been used for Data Visualization purpose. The machine hardware configuration includes Intel i7 dual core processors, 8 GB RAM, 4 GB NVIDIA 8700 GTX graphic card and 1 TB hard disk. The algorithm implementation and sentiment analysis has been performed in Java. For the purpose of data storage, MongoDB, a NoSQL database, was installed in Windows environment. Due to incompatibility of data visualization libraries on Windows environment, those particular libraries has been installed on Linux environment. Data visualization for Histograms and Time Series analysis has been implemented using *Plotly* libraries , whereas Sentiment Word Cloud has been implemented in Java using *Kumo* libraries inside Windows environment.

5.1 Histograms

In this section, I have presented histograms of five top most mentioned companies in Reuters during 10 year time period between 2008 and 2017. Each candlestick represents one month duration.

5.1.1 Morgan Stanley Capital International (MSCI)

MSCI Inc., is a global provider of equity, fixed income, hedge fund stock market indexes, and multi-asset portfolio analysis tools. This company falls under financial sector. This is most mentioned company with 59489 news article during the 10 year period.

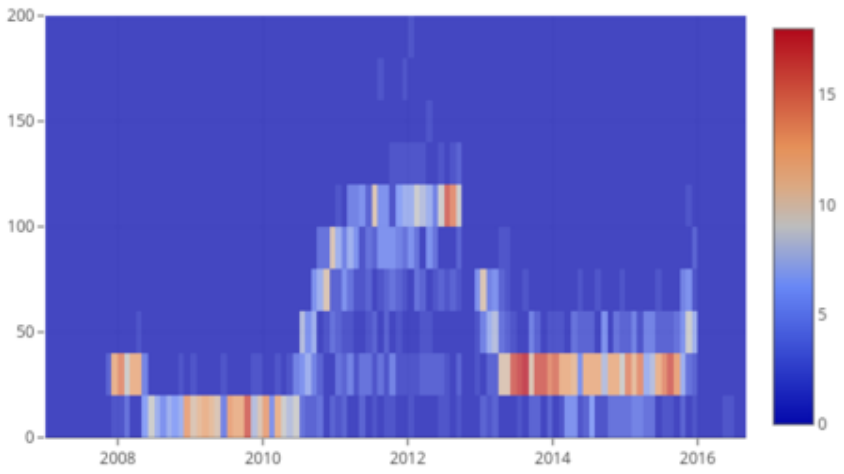

Figure 5.2: MSCI-Histogram

The news article primarily talks about different funds types launched by MSCI during this period. The launch of this funds can be directly correlated with the stock price.

5.1.2 Morgan Stanley

Morgan Stanley, the second most mentioned company in the same 10 year period, has appeared in news for 17,888 times. Morgan Stanley is an American multinational investment bank and financial services company.

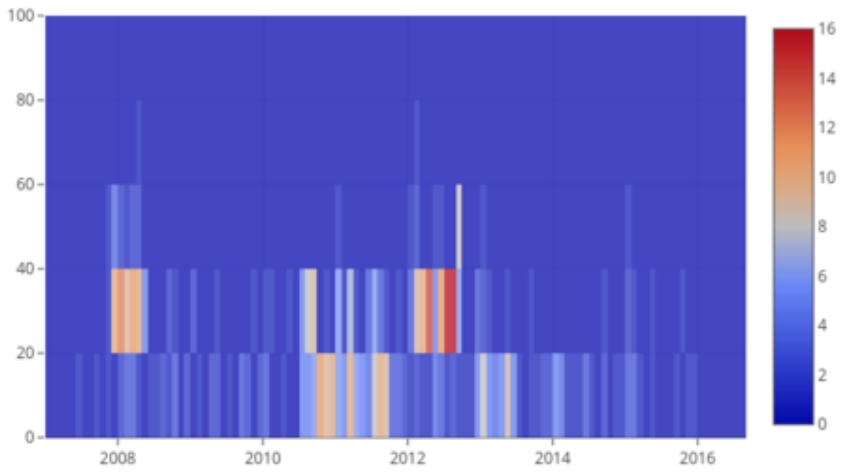

Figure 5.3: Morgan Stanley-Histogram

Although this company deals in financial sector, the major spikes in histograms are results of company's annual conference, known as Morgan Stanley Technology, Media Telecom Conference. This kind of news article are not directly correlated with company's actual performance.

5.1.3 Goldman Sachs

Goldman Sachs is another American multinational investment bank and financial service company. It offers services in investment management, securities, asset management, prime brokerage, and securities underwriting. It stands at third position in the list of most mentioned companies. It has been mentioned in the news 10046 times during the 10 year period.

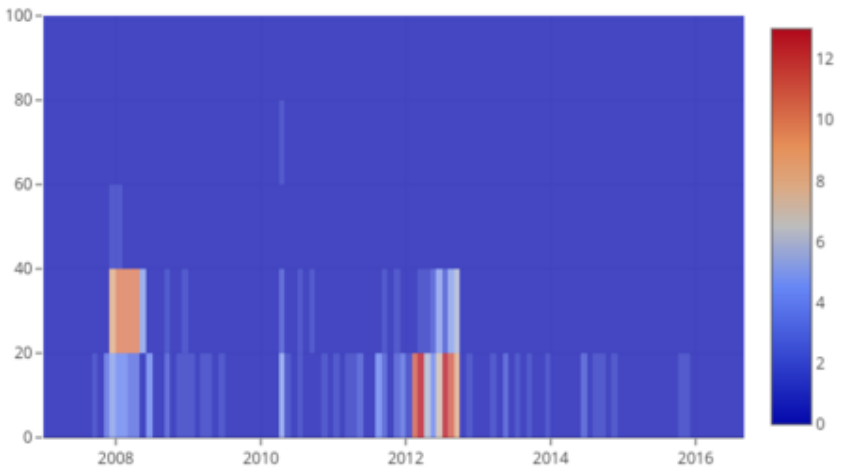

Figure 5.4: Goldman Sachs-Histogram

Major spike in news article can be seen during the end of first quarter of 2010, when the company was facing SEC investigation. This kind of news article can result in affecting the stock price in negative manner.

5.1.4 NASDAQ

The NASDAQ is an American over-the-counter stock exchange. It is the one of the largest stock exchange in the world by market capitalization. It has been in news frequently for the 10 year period with total number of news article reached up to 9979.

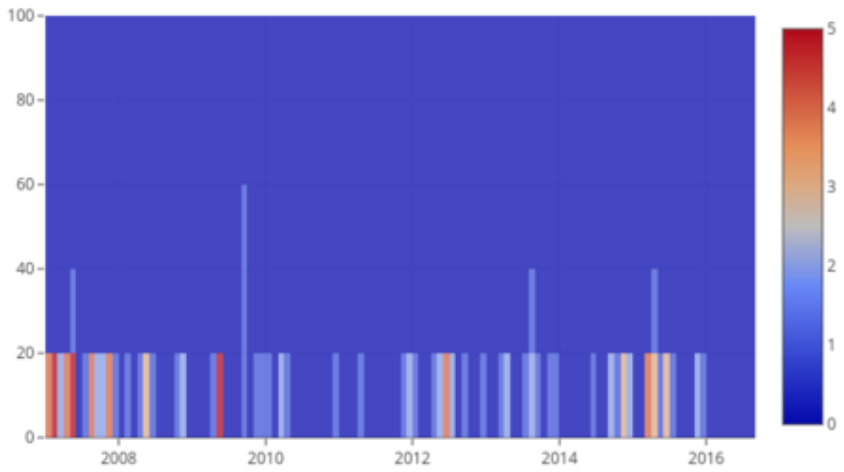

Figure 5.5: NASDAQ-Histogram

Although, NASDAQ stocks seen rising in the beginning of third quarter of 2009, it consists of similar kind of news articles. The news article published with respect to NASDAQ mainly talked about various kinds of notices issued to different companies. The nature of these notices is such that it is more concerned about the company rather than NASDAQ itself. This kind of news primarily affects the concerned stock.

5.1.5 Microsoft Corporation

Microsoft Corporation is an American multinational technology company that develops computer software, consumer electronics, personal computers, and related services.

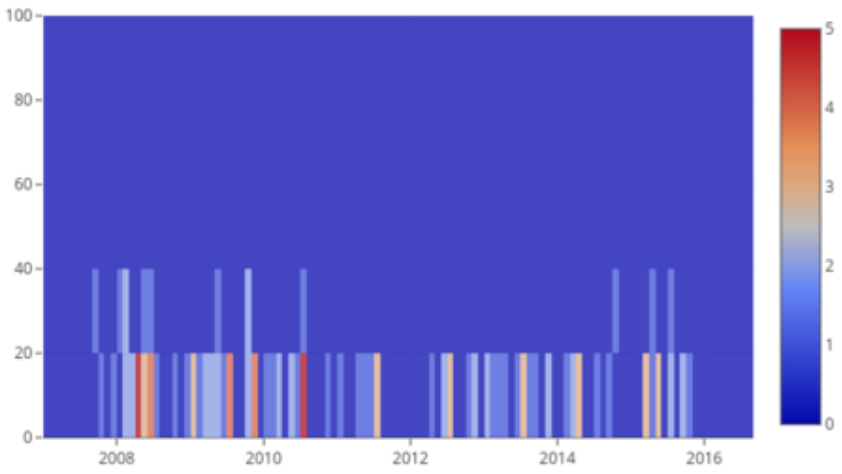

Figure 5.6: Microsoft-Histogram

Microsoft has been mentioned a total of 8712 times in the 10 year period. It is the only technology based company in the top 5 most mentioned companies in Reuters news article data set. The major spikes has been indicating events directly relating with company performances such as Microsoft $44.6 billion bid to buy Yahoo Inc. in early 2008 or its launch of corporate bond in mid 2015. This type of news can be directly correlated with the company's stock price.

Although, as mentioned before, Microsoft is the only technology based company in top 5 mentions, Apple Inc, a Microsoft rival in software domain has been mentioned almost equal number of times with 8556 during the same 10 year period.

5.2 Time Series Analysis

In this section, I present outcome of my analysis of direct relationship between news sentiment and company's stock price. The stock price has been obtained from *Yahoo! Finance* and sentiment score has been computed based on daily average of total sentiment score.

For experiments, I begin with analyzing the news article associated with the top mentioned companies during the 10 year period. After careful observation of top company news trends, I have decided to choose two technology based companies i.e. Apple and Microsoft, since their news article distribution is closely related to each other. In terms of news articles frequency, Apple and Microsoft share almost similar profile. Apple, albeit mentioned a bit lower number of times than Microsoft (8556 vs 8712),it has been mentioned on more days compare to Microsoft (2112 vs 2077). It this section, I would highlight two most mentioned news event of both companies and along with the impact of these news items on their stock price. In the table shown below, *NA* represent data Not Available for that particular day.

5.2.1 First Apple Watch launch

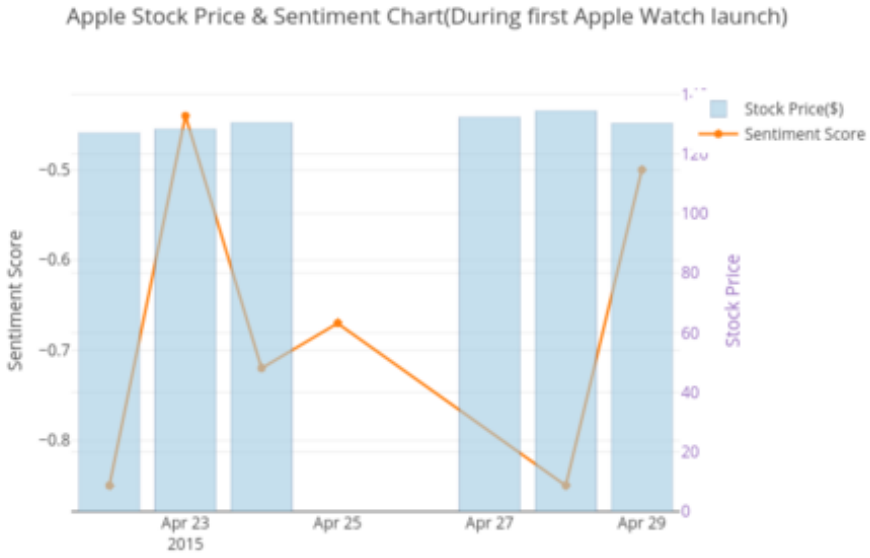

Figure 5.7: First Apple Watch launch

First Apple watch had launched on April 24, 2015. From figure 5.7 and Table 5.1, I have observed that the market sentiment and stock prices started improving before launch date itself due to positive anticipation by news media. These trend followed for couple of days in next week as well before price fell more than $4 in single day on April 29. Although market sentiment turned negative but their impact on stock prices were not visible. This observation shows that stock performance was not correlating with market sentiment.

Date	Stock Price($)	Average Sentiment Score	Total number of news article
2015-04-22	126.99	-0.85	7
2015-04-23	128.3	-0.44	9
2015-04-24	130.49	-0.72	22
2015-04-25	NA	-0.67	49
2015-04-26	NA	NA	NA
2015-04-27	132.31	NA	NA
2015-04-28	134.46	-0.85	14
2015-04-29	130.16	-0.5	14

Table 5.1: Apple Watch launch

5.2.2 Apple Music launch

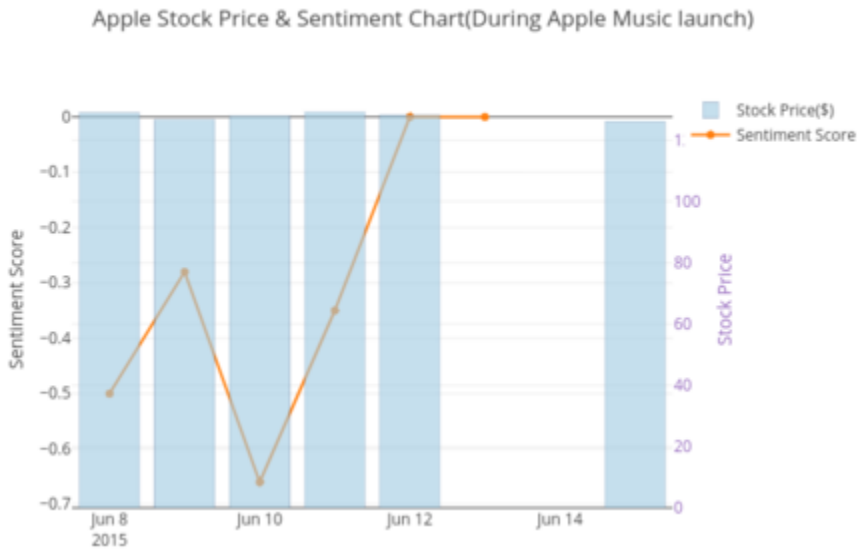

Figure 5.8: Apple Music launch

Another popular event in history of Apple in this 10 year time period was the launch of Apple Music. Although, Apple Music was launched on June 30, 2015, the news had been broken out on June 9. However, both news and stock performance

Date	Stock Price($)	Average Sentiment Score	Total number of message
2015-06-08	128.9	-0.5	2
2015-06-09	126.7	-0.28	32
2015-06-10	127.92	-0.66	6
2015-06-11	129.18	-0.35	14
2015-06-12	128.19	0	5
2015-06-13	NA	0	2
2015-06-14	NA	NA	NA
2015-06-15	126.1	NA	20

Table 5.2: Apple Music launch

only short lived for couple of days before stock price declined again and news has diverted to other topics, as can be observed from Figure 5.8 and Table 5.2. Based on this observation, I can conclude that a very short term positive relationship exists between stock prices and news sentiment.

5.2.3 Microsoft bid for Yahoo!

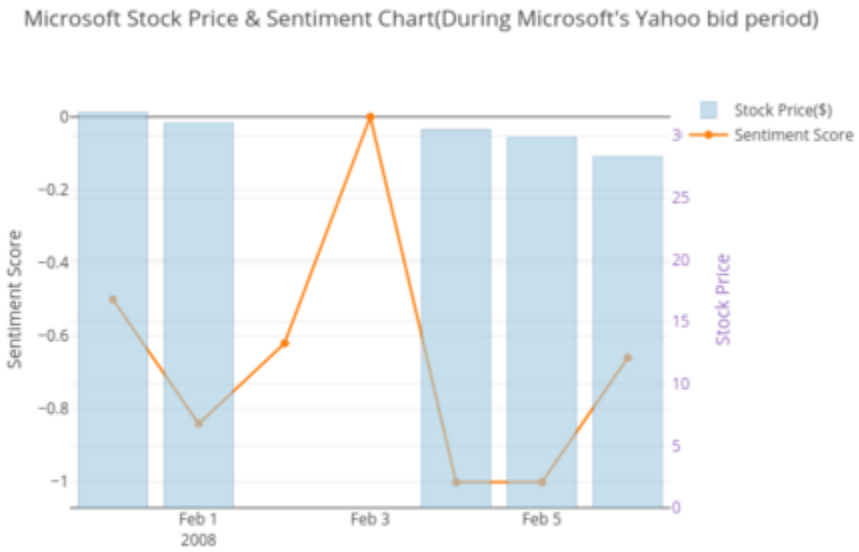

Figure 5.9: Microsoft bid for Yahoo

In case of Microsoft's bid for Yahoo!, both market and news sentiment reacted negatively towards the news. Although there was decline in number of news article but the stock kept plummeting in next week as well. Here (Figure 5.9 and Table 5.3), close relationship is visible between stock prices and news sentiment.

Date	Stock Price($)	Average Sentiment Score	Total number of message
2008-01-31	31.91	-0.5	2
2008-02-01	31.06	-0.84	26
2008-02-02	NA	-0.62	8
2008-02-03	NA	0	2
2008-02-04	30.49	-1	1
2008-02-05	29.91	NA	NA
2008-02-06	28.34	-1	1

Table 5.3: Microsoft bid for Yahoo!

5.2.4 Microsoft Corporate bond launch

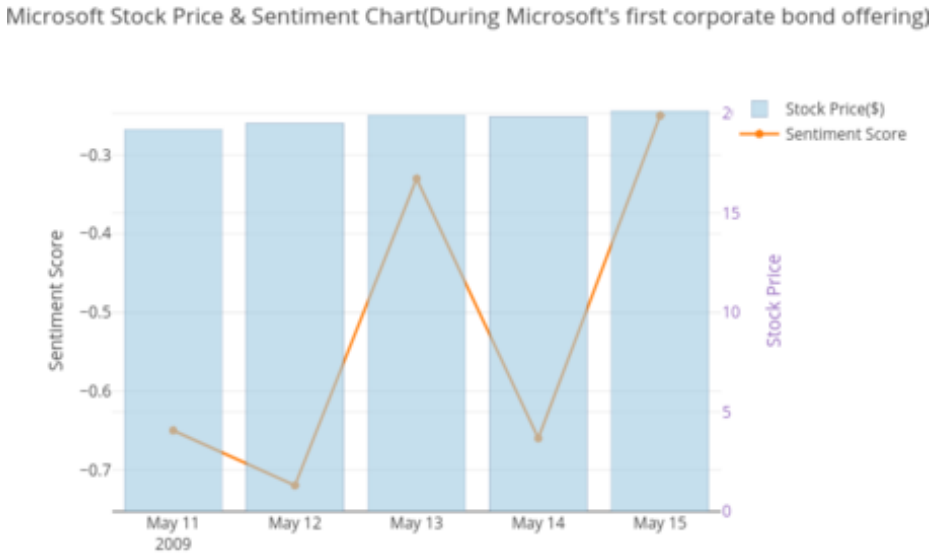

Figure 5.10: Microsoft Corporate bond launch

On the contrary to previous cases, in this case of Microsoft's bond launch, market has reacted positively to news for next few days (Figure 5.10 and Table 5.4),

Date	Stock Price($)	Average Sentiment Score	Total number of message
2009-05-11	19.2	-0.65	23
2009-05-12	19.51	-0.72	11
2009-05-13	19.52	-0.33	3
2009-05-14	19.83	-0.66	3
2009-05-15	20.13	-0.25	4

Table 5.4: Microsoft Corporate bond launch

whereas news sentiments were divided for the same time period. As it is clear from Figure 5.10 that there exist no concrete relationship between stock price and news sentiment.

As it is evident from above figures and tables that after the release of news the stocks prices went high up to a week with the exception of Microsoft's bid for Yahoo!. In that particular case, the news sentiment were negative as well. As reported in news the next day, even Yahoo!, was also cautious of this offer. If we observe the trends for longer period we can see that relationship between news sentiment and stock prices is diminishing. One of the reasons for this can be attributed to the fact that these events/news have been mixed with other news items, which belong to same company but reporting other incidents. Although, this news item can effect the sentiment score but this news items are completely eclipsed in front of major product's launch. Another factor to note is that the total number of news articles per day. This factor was not shown in time series chart but mentioned in table. More number of news articles, as they are on the same day or next, can be attributed to cause more impact given the widespread use of social media.

5.3 Sentiment Word Cloud

I have used *Kumo Java library* for constructing sentiment word cloud. The green and red colors in the word cloud are representing the positive and negative sentiments respectively. The word frequency is directly proportional to word size and opacity. During experiments, I have not found any relationship between word placements and relative placements between positive and negative sentiment word cloud. One important fact to mention is that the word cloud narrows down the keywords to substantial level but the requirement of careful observation still remains by professional expert, to gain useful insight. In this section, I have shown the top five mentioned stocks in news media.

5.3.1 Morgan Stanley Capital International (MSCI)

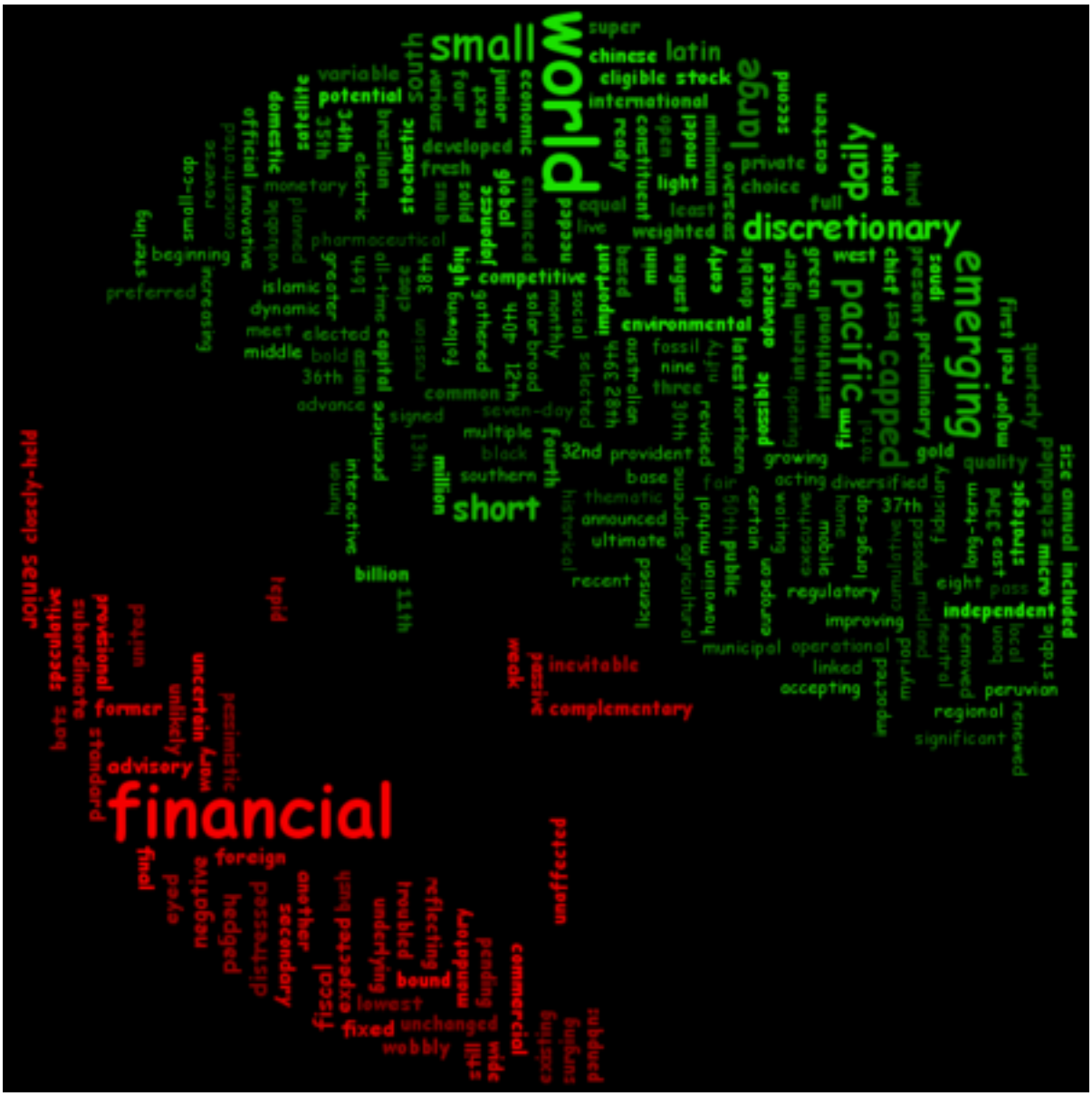

Figure 5.11: MSCI Sentiment Word Cloud

5.3.2 Morgan Stanley

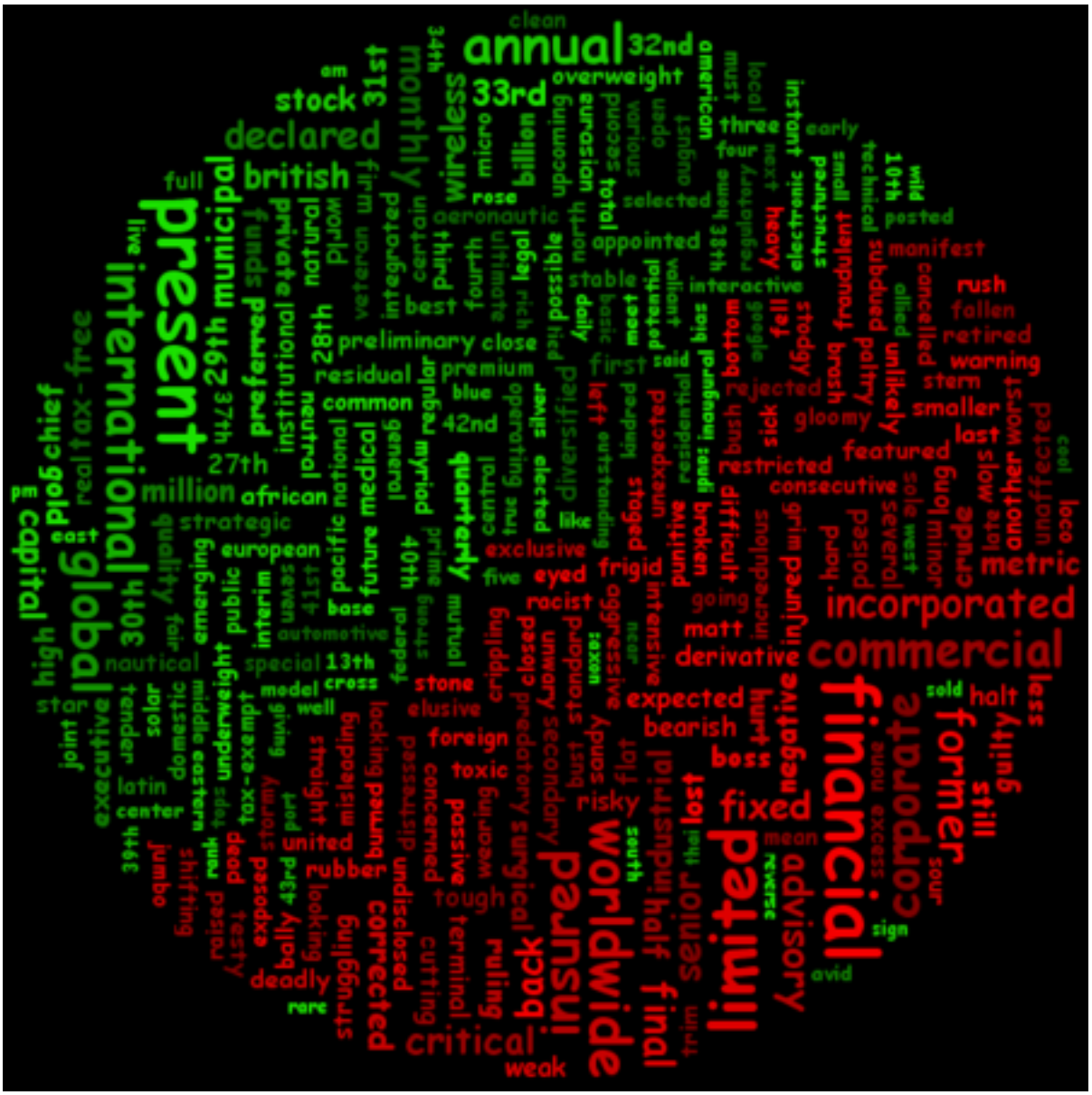

Figure 5.12: Morgan Stanley Sentiment Word Cloud

5.3.3 Goldman Sachs

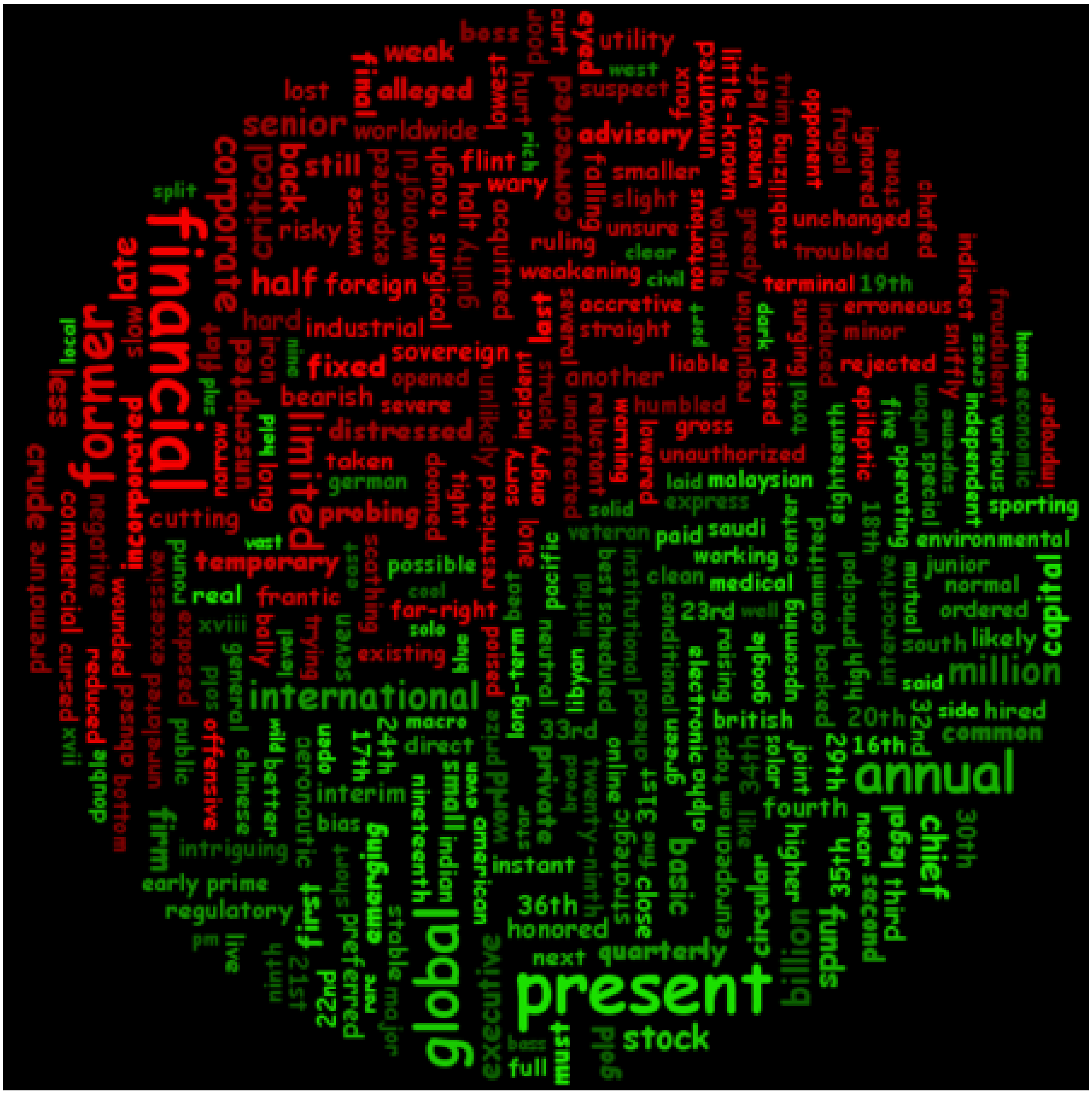

Figure 5.13: Goldman Sachs Sentiment Word Cloud

5.3.4 NASDAQ

Figure 5.14: NASDAQ Sentiment Word Cloud

5.3.5 Microsoft Corporation

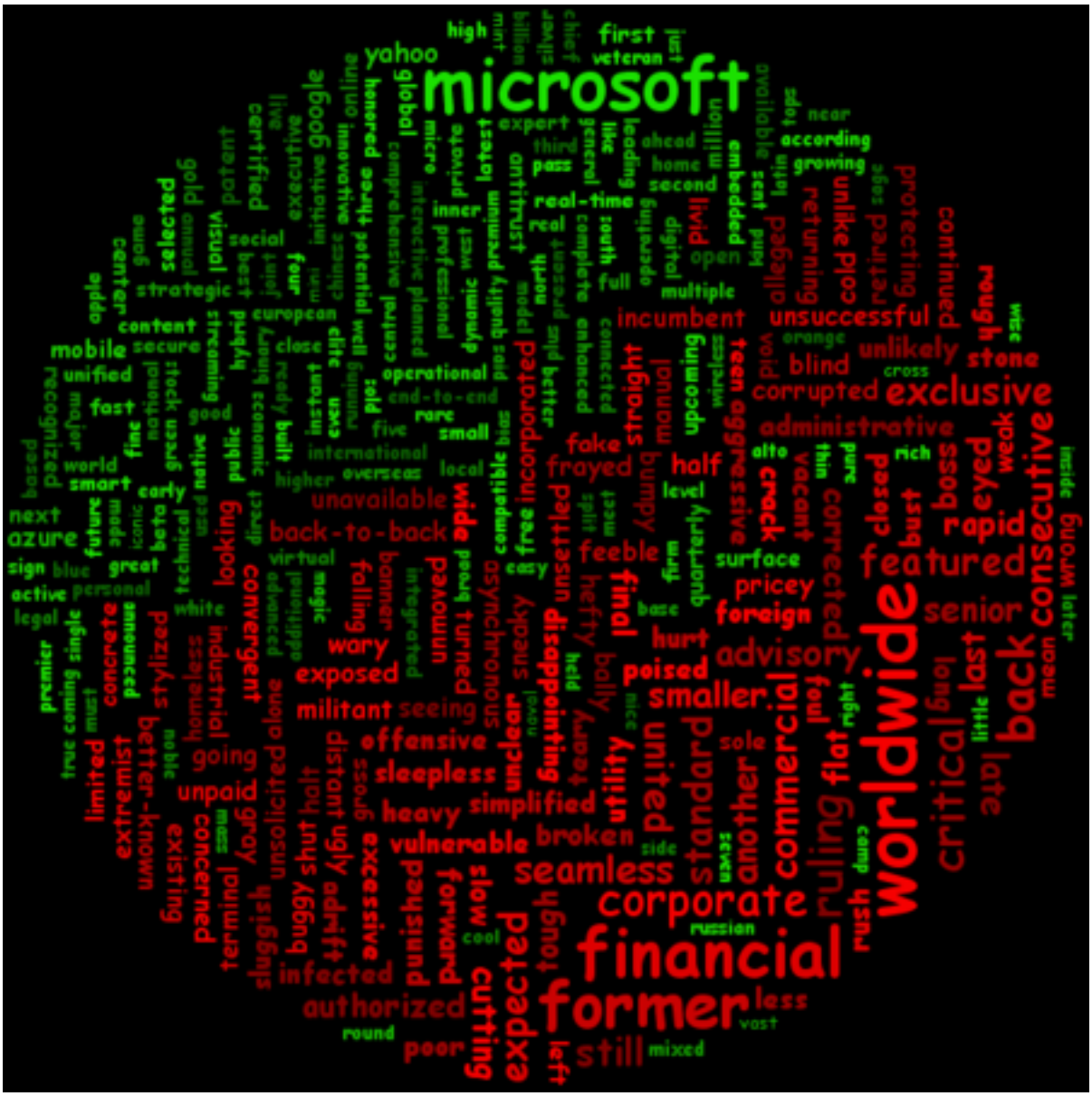

Figure 5.15: Microsoft Corporation Sentiment Word Cloud

At the first glance, it seems obvious that each company word cloud is representing keywords from its business sector such as in financial sector company's

word cloud filled with key word like 'finance' or finance related keywords. In case of financial sector, for example, MSCI, Morgan Stanley, Goldman Sachs and NASDAQ, the word cloud seems generic and it only provides information at a shallow level. However, there are few popular keywords also present, which can affect company performance. For example,in the case of Microsoft, we can observe that 'google' and 'yahoo' are present as well. The 'Yahoo!' keyword trend has been also observed while performing Histogram experiments, where finding shows the news regarding Microsoft buying Yahoo!. In Microsoft word cloud, we can also find mention of it's product such as 'azure' and 'xbox'. These product and service review can also affect Microsoft stock price.

Overall, these sentiment word clouds has provided the most significant word mentioned with regards to these specific companies. In terms of major findings, we can observe that trends associated with the companies. This trends may or may not have effect on company stock prices but it can still give an idea about the most mentioned events associated with the company.

Chapter 6

Conclusion

During research, one important trend I observed is that data or information whether it is coming from social media or news media has become one of the most expensive commodity in itself. The scraping of news media site become harder in recent years. In case of social media, like Facebook or Twitter getting data using their APIs have become lot more expensive as compared to past years. Combining cost factor with the explosion of data generation made data gathering highly expensive. In this thesis, I have used Reuters news article data set to find relationship between news sentiments and stock prices.

In this research, I have used three different data visualization techniques, Histogram, Time Series series analysis and Sentiment Word cloud for the purpose of analyzing data distribution, discovering relationship between news sentiment and stock price, and finding other trends affecting the market sentiment for the company respectively. First, in case of histogram, I have found that data distribution of news article started reducing immensely in the case of top five mentioned companies itself.

However, this problem can be reduced by analyzing two or more companies in same time period. Second, in the case of time series analysis, the relationship between news media sentiments and stock prices are loosely identical. The main reason behind this observation can be attributed to news article that went unobserved by market but picked up during sentiment analysis. Last, in sentiment word cloud, I have displayed most frequent appearing words with their sentiment polarity. These sentiment word clouds substantially backed up the patterns discovered in previous data visualization experiments.

During this research I have found certain degree of relationship between news media sentiments and stock prices, where it can be used for profit making in real market. This study is by no means exhaustive, primary reason behind this shortcoming can be attributed to lack of source of information as compared to bigger firms. This research can be further expanded by acquiring and studying larger source of information.

Bibliography

[1] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," *arXiv preprint arXiv:1703.03130*, 2017.

[2] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank," in *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013, pp. 1631–1642.

[3] A. Radford, R. Jozefowicz, and I. Sutskever, "Learning to generate reviews and discovering sentiment," *arXiv preprint arXiv:1704.01444*, 2017.

[4] Bloomberg Markets Magazine, "Finding novel ways to trade on sentiment data," https://www.bloomberg.com/professional/blog/finding-novel-ways-trade-sentiment-data/, 2017, [Online; accessed 15-July-2018].

[5] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of computational science*, vol. 2, no. 1, pp. 1–8, 2011.

[6] jackp. montreal-and-san-francisco-temperatures. https://plot.ly/ jackp/10010 [Last accessed: April 21, 2019].

[7] A. Rasool, R. Tao, K. Marjan, and T. Naveed, "Twitter Sentiment Analysis: A Case Study for Apparel Brands," in *Journal of Physics: Conference Series*, vol. 1176, no. 2. IOP Publishing, 2019, p. 022015.

[8] Aatif Sulleyman, "Twitter introduces 280 characters to all users," https://www.independent.co.uk/life-style/gadgets-and-tech/news/twitter-280-characters-tweets-start-when-get-latest-a8042716.html, 2018, [Online; accessed 23-Aug-2018].

[9] Will Nicol, "What is reddit? a beginners guide to the front page of the internet," https://www.digitaltrends.com/social-media/what-is-reddit/, 2018, [Online; accessed 23-Aug-2018].

[10] StockTwits, Inc., "Introducing the StockTwits Cashtag Collection," https://blog.stocktwits.com/introducing-the-stocktwits-cashtag-collection-59d399fbf796, 2016, [Online; accessed 23-Aug-2018].

[11] Statista, "Most famous Social Network sites worldwide as of july 2018, ranked by number of active users (in millions)," https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/, 2018, [Online; accessed 23-Aug-2018].

[12] Alana Abramson, "Facebook says Russian Accounts Bought $100,000 in

Ads during the 2016 Election," http://time.com/4930532/facebook-russian-accounts-2016-election/, 2017, [Online; accessed 23-Aug-2018].

[13] Rachael Revesz, "Toyota loses $1.2bn in value five minutes after Donald Trump's tweet," https://www.independent.co.uk/news/world/americas/toyota-12bn-value-plummet-shares-stock-market-donald-trump-tweet-move-mexico-tax-a7512096.html, 2017, [Online; accessed 16-July-2018].

[14] Kaya Yurieff, "Snapchat stock loses $1.3 billion after Kylie Jenner tweet," https://money.cnn.com/2018/02/22/technology/snapchat-update-kylie-jenner/index.html, 2018, [Online; accessed 16-July-2018].

[15] Ana Alexandre, "John Mcafee charges $105,000 per tweet for promoting cryptocurrency projects," https://cointelegraph.com/news/john-mcafee-charges-105000-pertweet-for-promoting-cryptocurrency-projects, 2018, [Online; accessed 17-July-2018].

[16] "Award-winning Audi Twitter strategy," https://www.catalystdigital.com/digital-marketing-case-studies/audi-twitter-strategy/, [Online; accessed 10-December-2018].

[17] T. H. Nguyen, K. Shirai, and J. Velcin, "Sentiment Analysis on Social Media for stock movement prediction," *Expert Systems with Applications*, vol. 42, no. 24, pp. 9603–9611, 2015.

[18] A. Mittal and A. Goel, "Stock prediction using Twitter sentiment analysis," *Standford University, CS229 (2011 http://cs229. stanford.*

edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf), vol. 15, 2012.

[19] W. Chen, K. Lai, and Y. Cai, "Topic generation for Chinese stocks: a cognitively motivated topic modeling method using social media data," *Quantitative Finance and Economics*, vol. 2, no. 2, pp. 279–293, 2018.

[20] K. Daniel, D. Hirshleifer, and A. Subrahmanyam, "Investor psychology and security market under-and overreactions," *the Journal of Finance*, vol. 53, no. 6, pp. 1839–1885, 1998.

[21] E. F. Fama, "Market efficiency, long-term returns, and behavioral finance," *Journal of financial economics*, vol. 49, no. 3, pp. 283–306, 1998.

[22] H. Hong and J. C. Stein, "A unified theory of underreaction, momentum trading, and overreaction in asset markets," *The Journal of finance*, vol. 54, no. 6, pp. 2143–2184, 1999.

[23] R. Stagner, "The Cross-Out Technique as a Method in Public Opinion Analysis," *The Journal of Social Psychology*, vol. 11, no. 1, pp. 79–90, 1940. [Online]. Available: https://doi.org/10.1080/00224545.1940.9918734

[24] S. A. Sandri, D. Dubois, and H. W. Kalfsbeek, "Elicitation, assessment, and pooling of expert judgments using possibility theory," *IEEE transactions on fuzzy systems*, vol. 3, no. 3, pp. 313–335, 1995.

[25] M. V. Mäntylä, D. Graziotin, and M. Kuutila, "The evolution of sentiment

analysisa review of research topics, venues, and top cited papers," *Computer Science Review*, vol. 27, pp. 16–32, 2018.

[26] P. R. Wheale and L. H. Amin, "Bursting the dot.com "bubble': A Case Study in Investor Behaviour," *Technology Analysis & Strategic Management*, vol. 15, no. 1, pp. 117–136, 2003. [Online]. Available: https://doi.org/10.1080/0953732032000046097

[27] Vanessa Page, "Warren Buffett: Most influential quotes," https://www.investopedia.com/university/warren-buffett-biography/warren-buffett-most-influential-quotes.asp, [Online; accessed 9-December-2018].

[28] C. d. Santos, M. Tan, B. Xiang, and B. Zhou, "Attentive pooling networks," *arXiv preprint arXiv:1602.03609*, 2016.

[29] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," *arXiv preprint arXiv:1503.00075*, 2015.

[30] M. Qasem, R. Thulasiram, and P. Thulasiram, "Twitter sentiment classification using machine learning techniques for stock markets," in *Advances in Computing, Communications and Informatics (ICACCI), 2015 International Conference on.* IEEE, 2015, pp. 834–840.

[31] P. Houlihan and G. G. Creamer, "Can Sentiment Analysis and Options Volume Anticipate Future Returns?" *Computational Economics*, vol. 50, no. 4, pp. 669–685, 2017.

[32] A. Ceron, L. Curini, S. M. Iacus, and G. Porro, "Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens political preferences with an application to italy and france," *New Media & Society*, vol. 16, no. 2, pp. 340–358, 2014.

[33] Bloomberg Markets Magazine, "Trading the news: Use machine-readable data to find alpha," https://www.bloomberg.com/professional/blog/trading-news-use-machine-readable-data-find-alpha/, 2016, [Online; accessed 15-July-2018].

[34] Z. Chen and A. Lu, "Slow diffusion of information and price momentum in stocks: Evidence from options markets," *Journal of Banking & Finance*, vol. 75, pp. 98–108, 2017.

[35] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins, "The predictive power of online chatter," in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining.* ACM, 2005, pp. 78–87.

[36] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proceedings of the 42nd annual meeting on Association for Computational Linguistics.* Association for Computational Linguistics, 2004, p. 271.

[37] X. Yu, X. Xin, L. Chen, and H. S. Kim, "Predicting Market Reactions to Bad News," 2018.

[38] Twitter, "Rate Limiting," https://developer.twitter.com/en/docs/basics/rate-limiting.html, [Online; accessed 16-November-2018].

[39] David Wallach, "Financial Tweets," https://www.kaggle.com/davidwallach/financial-tweets/home, [Online; accessed 10-December-2018].