Public Sentiment on Twitter and Stock Performance

A Study in Natural Language Processing

Jimmy Henriksson and Carl Hultberg

Abstract

Since recent years, the use of non-traditional data sources by hedge funds in order to support investment decisions has increased. One of the data sources which has increased most is social media and it has become popular to analyze the public opinion with help of sentiment analysis in order to predict the performance of a company.

In order to evaluate the public opinion one need big sets of Twitter data. The Twitter data was collected by streaming the Twitter feed and the stock data was collected from a Bloomberg Terminal. The aim of this study was to examine if there is a correlation between the public opinion of a stock and the stock price, and also what affects this relationship. While such a relationship cannot be established in general, we are able to show that if the data quality is good, there is a high correlation between the public opinion and stock price, and that significant events surrounding the company results in a higher correlation during that period.

Abstract

Dem senaste åren har användandet av icke-traditionella datakällor ökat av hedgefonder för att ta investeringsbeslut. En av datakällorna som blivit populära är sociala medier och det har blivit vanligt att analysera folkopinionen med hjälp av sentimentanalys för att kunna förutspå ett företags resultat.

För att analysera folkopinionen krävdes stora mängder Twitter-data. Twitter-datan hämtades genom att strömma Twitter-flödet och aktiedatan hämtades från en Bloomberg Terminal. Målet med studien var att undersöka ifall det finns en korrelation mellan folkopinionen av en aktie och aktiens prisutveckling, och även vad som påverkar denna relationen. Även om en sådan relation inte kan fastställas i allmänhet så kan vi visa att om datakvaliten är god, så finns det en hög korrelation mellan folkopinionen och aktiepriset, samt att vid betydande händelser som rör företaget, så resultar det i en hög korrelation under den perioden.

Key Words

Sentiment Analysis; Public Opinion; Stock Market; Stock Performance; Statistics; Computer Linguistics; Natural Language Processing; Twitter.

Nyckelord

Sentimentanalys; Folkopinion; Aktiemarknaden; Aktiekursutveckling; Statistik; Språkteknologi; Twitter.

Acknowledgements

We want to thank our supervisor Arvind Kumar for his very valuable guidance during the process of writing this thesis.

Swedish Title

Allmäna Sentimentet på Twitter och Aktiemarknaden
- en studie i språkteknologi

Authors

Jimmy Henriksson and Carl Hultberg
School of Electrical Engineering and Computer Science
KTH Royal Institute of Technology

Place for Project

Stockholm, Sweden

Examiner

Örjan Ekeberg
School of Electrical Engineering and Computer Science
KTH Royal Institute of Technology

Supervisor

Arvind Kumar
School of Electrical Engineering and Computer Science
KTH Royal Institute of Technology

Contents

1 Introduction

Since recent years the use of social media has increased and today more than three billion people use some social medium daily and in The United States 25% of all adults use twitter [23]. Social media messages are believed to be the largest data source for public opinion [22]. Hedge funds have during the last years started to broaden their analysis by using non-traditional data sources such as social media when making investment decisions [13]. According to a survey conducted 2017 by Ernst & Young 27% of the hedge funds said that they have used social media. By comparison 25% used credit card data, 14% used satellite images and 10% tracked installation of applications [18]. The main use of this data has been to create a sentiment of companies which is useful for a long-term Long/Short strategy [20].

In order to quantify emotional content expressed in large amounts of text data, automatic sentiment analysis can be used, which this thesis does. Sentiment analysis refers to the process of automatically detecting if a text contains emotional or opinionated content and determine its polarity. This makes it possible to efficiently investigate the public opinion about different stocks.

Many investors use social media sentiment as a parameter when making investment decisions of stocks [22]. Therefore, it is interesting for investors to know how strong the correlation is between the social media sentiment and different stocks price performance and what factors affects this potential relationship.

1.1 Purpose

The purpose of the thesis is to show the significance and correlation of the sentiment on social media when it comes to the price performance of stocks in different sectors. Based on this study, investors can choose a more reasonable weight of this parameter when making an investment, which potentially can lead to a higher return. Investors can also get an idea of when it is more appropriate to use social sentiment when making investment decisions, and whether it suits some companies or sectors better than others. The study strives to get a closer understanding of how and when to leverage this information.

Previous research has focused on stock prediction using social sentiment rather than examining the correlation and how and when it is more suitable to use the sentiment, and if any other relationships exists.

1.2 Research Question

This study aims to gather Twitter sentiment data regarding particular stocks in different sectors in order to analyze how strong the correlation is between the Twitter sentiment and the stock performance. This leads to the following research question: *Is there a correlation between Twitter Sentiment of stocks and its performance, and what factors affect this relationship?*

1.3 Scope

In this study we have for practical reasons, such as the limited time interval and relevant data, chosen only to analyze data from the social media Twitter during a six week time interval. Furthermore, the analysis will not value more levels of sentiment than the sentiment score interval -1 to 1. The thesis focuses only on a limited selection of US stocks.

2 Theoretical Background

This chapter explains the following key concepts of the report:

- Natural Language Processing

- Sentiment Analysis

- Twitter

- The Stock Market

- Correlation and Causality

Lastly, the chapter presents a section on previous research related to the thesis.

2.1 Natural Language Processing

Natural Language Processing (NLP) is a field in computer science which involves the interaction between humans natural language and computers, and how to analyze and process the human language data. This is a tough challenge since computers often requires us to talk to them with a programming language which is unambiguous and structured. Natural human language is ambiguous and its structure is dependent on many variables such as language, slang and social context.

2.2 Sentiment Analysis

Sentiment analysis refers to the process of automatically detecting if a text contains emotional or opinionated content and determine its polarity [12]. Sentiment analysis of social media is used to get an idea of the public mass' views on different topics. It has been used to predict presidential elections, as the twitter sentiment correlates with the public masses' views [26] and also what the public's perception of different stocks is in order to predict the performance [16]

2.2.1 Amplifiers

An amplifier, also known as an intensifier, is a word which increases the polarity. An example is "I *really* like Amazon". An amplifier increases the polarity by 1.8 with the

standard weighting.

2.2.2 De-amplifiers

A de-amplifiers is the opposite of an amplifier, e.g a down-toner which decreases the polarity. An example is "I *hardly* like Amazon". A de-amplifier decreases the polarity by 0.8 with the standard weighting.

2.2.3 Negators

A negator is commonly used in our day to day communication and an example is "I do *not* like it". If the sentence "I like it" would get the score 1 then "I do not like it" would get -1 since it flips the score. To be more precise, an amplifier becomes a de-amplifier if it contains an odd number of negators and vice versa.

2.2.4 Adversative Conjuction

Adversative conjuction is when you use, for example, "but", "however" or "although" in a sentence. An adversative conjuction before a polarized word increases the polarity of the word and if it is after a polarized word it decreases the polarity. This is due to the belief that an adversative conjuction makes the next part of the sentence more important than the part before. An example of this is "The new Apple product looks great, *but* the overall impression is bad"

2.2.5 Accuracy of Sentiment Analysis

In its current stage, automated sentiment analysis is not able to be as accurate as if it were performed by a human. Basic sentiment analysis is performed by classifying each word in a sentence separately by determining how negative or positive each word is, and how they affect each other. A text segment consisting of two words, where the first word is classified as negative and the second word strongly positive, can be classified as highly positive by a human being. But if these words are weighed together without checking how they affect each other, it ends up with a value that is close to neutral. Understanding how words affect each other is one of the challenges in the area. Another challenge

when it comes to analyzing the sentiment on social media is that the language is often informal and consists of slang words and emoji-symbols. Previous studies have shown that 97 percent of all comments on the social media website Myspace contain at least one non-standard language feature [24].

2.3 Twitter

Twitter is an online microblogging platform that allows users to broadcast short posts no longer than 140 characters, called tweets. A user can broadcast tweets, follow other users and choose to publish the tweets publicly or within its private network. Follow a user means that this user's tweets will get shown on the user's feed. The follow function is not mutual. The platform had 321 million active users per month in 2018 and about 500 million tweets per day. Twitter uses hashtags to help users broadcast and find tweets that they are interested in. A hashtag is a "#"-sign followed by a keyword. These keywords can be anything except spaces. In this way, users can find each other and discuss specific subjects [27].

2.3.1 Twitter Standard Search API

Twitter offers developers a search API where the user can search for tweets. The API helps the user to search for tweets that contains selected words at specific dates. The limitations with the API is that the user only can search for tweets up to seven days back in time, and that the tool do not support data fetching by using complex queries. The API is also limited to fetch not more than 18.000 tweets every fifteen minutes. Twitter explains on the developer page that the tool is focused on relevance and not completeness. This may result in some tweets and users may be missing from the search result [28].

2.4 The Stock Market

Many corporations start out as small businesses with limited capital. A common way to raise capital is by selling shares of the business to investors in order to expand. These shares of the company are referred to as stocks. Investors buy a stock with the purpose of making money through dividends or that the stock price increases. A stock exchange is a

marketplace where an investor can buy and sell securities like stocks. The stock exchange works like a auction house [15].

Supply and demand are the two factors that determine the price of a stock. Supply and demand in turn depends on fundamental factors, technical factors and market sentiment. Examples of fundamental factors are the perceived risk of the stock, discount rate and expected growth. Technical factors includes inflation, economic strength of the market, trends and the stocks liquidity. Market sentiment refers to the psychology of stock market participants, how the participants interpret the current state of the stock price [14].

2.4.1 Efficient-Market Hypothesis

Stock market prediction has been on the table for decades and has brought attention both from business and in science. One of the first theories on the subject was the Efficient-Market Hypothesis (EMH) which states that asset pricing reflects all available information. According to the EMH assets always trade on a fair value and markets only reacts on new information such as news, therefore is it impossible to beat the market over time on a risk-adjusted basis. News are unpredictable and this suggest that asset pricing follows a random walk with a 50% chance of going either up or down [17]. EMH and The Random Walk Theory has received a lot of criticism and evidence show that asset pricing do not follow a random walk and in other words asset pricing can be predicted to some degree. For example, Warren Buffett wrote a article arguing that if several funds managed to beat the index year after year it cannot be an random event [8]. The NYU Stern School of Business professor Aswath Damodaran referred this as a proof that the market is not always efficient [9].

2.5 Correlation and Causality

Correlation is commonly defined as how close two variables are to having a linear relationship with each other. The co-variance of two events X and Y is defined below and $E[X]$ is the expected value:

$$C(X,Y) = E[(X - E[X])(Y - E[Y])]$$

If $C(X,Y) = 0$ then they are said to be uncorrelated [7].

The correlation is defined as:

$$\rho(X,Y) = \frac{C(X,Y)}{D(X)D(Y)}, C(X,Y) \in [-1,1]$$

where $D(X)$ and $D(Y)$ represent the standard deviation of X and Y.

If $\rho(X,Y) = 1$ there is a perfect positive relationship between X and Y and if X increases Y also increases.

If $\rho(X,Y) = -1$ there is a perfect negative relationship between X and Y and if X increases Y decreases.

Even-though the correlation coefficient is -1 or +1 it does not have to imply that one causes the other. Causality refers to when one thing, often referred to the cause, gives rise to another cause which is often called the effect. It is important to differentiate association with causation, association refers to event which happen together more regularly than others, but that does not mean that the relationship is meaningful. In this thesis we will look at the correlation between public sentiment and stock price in general and if any of the two causes the other in particular.

2.6   Related Work

Recent studies show that investors can get early indicators of news by scraping different data sources such as blogs, social media and search engines. For example, Google used their software called Google Trends and showed that Google search queries could be used in order to predict consumer spending [25].

A previous paper by Bollen, Mao and Zeng show that mood and emotions are critical in financial decision making, and therefore it is safe to assume that mood and emotions also may impact asset pricing [20].

Mao, Wang, Wei and Liu found that there was a significant correlation between the daily number of tweets that mentions stocks in the stock index Standard & Poor 500 and stock indicators that are related to the price and volume. They also found that there was a significant correlation between the daily number of tweets and the daily traded volume

for eight out of ten industry sectors [19].

Another study found that there is a correlation between the sentiment on Twitter and the price of Microsoft's stock [21].

3 Methods

In this chapter the method used in the research and data collection is described. Furthermore, a thorough description of methods for sentiment analysis and data cleaning are described.

3.1 Data Collection

To be able to perform the sentiment analysis and calculate the correlation, Twitter and stock data sets is necessary. The collection of these data sets is described below.

3.1.1 Stock Data Collection

The stock data has been collected from a Bloomberg Terminal, exported to a csv-file. The data includes the close prices of the stocks during the analyzed time period. An example of the collected data from the Apple stock is presented in table 3.1 below.

AAPL US Equity	
Dates	PX_LAST
2019-03-04	175,85
2019-03-05	175,53
2019-03-06	174,52
2019-03-07	172,5
2019-03-08	172,91
2019-03-11	178,9
2019-03-12	180,91
2019-03-13	181,71

Table 3.1: Sample of the collected Apple stock data.

In order to ensure a varied set of companies in different industries, the companies presented below were selected.

American International Group   AIG is a global insurance company with operations in more than 80 countries and jurisdictions. They provide a range of insurance products

in business and in life, including: life insurance, general property and financial services. AIG's industry is Insurance [6].

Apple Inc. Apple designs and manufactures electronics such as phones, computers and smart watches and sells a range of related software. Apple's segments include the Americas, Europe, Japan, Greater China and Rest of Asia Pacific. Apples industry is Consumer Electronics [1].

The Boeing Company Boeing is the world's largest aerospace company. They are a manufacturer of commercial jetliners, defense, space and security systems, and service provider of aftermarket support. Boeing supports airlines and U.S. and allied government customers in more than 150 countries. Boeing's industry is Aerospace & Defense [2].

Facebook Inc. Facebook is an online social media and social networking service company. They are building products that enable people to connect through personal computers, mobile devices and other surfaces. Facebook's industry is Internet Content & Information [3].

Netflix Inc. Netflix is an internet entertainment service and are operative in more than 190 countries. Customers can watch TV series, documentaries and feature films across a wide variety of genres and languages on any internet-connected screen. Netflix's industry is Media [5].

Home Depot Inc. Home Depot is a home improvement retailer. They sells an assortment of building materials, lawn and garden products, and home improvement products. Home Depot's industry is Home Improvement Stores [4].

3.1.2 Twitter Data Collection

The Twitter data has been collected by using the Twitter Standard Search API, together with the programming language R. Since the API is limited to fetch not more than 18.000 tweets every fifteen minutes up to seven days back in time, an implementation of a script that handled the job automatically was necessary. The script was running every day for

6 weeks and collected 408.464 tweets posted between 4th March and 15th April 2019. The data was stored as different R data frames grouped by dates and stocks. To collect tweets related to the stocks, the script searches for the tweets that contain a dollar sign followed by the stock ticker for each company. All of the keywords used for the Twitter data collection are attached in the appendix.

R Programming Language R has been used for collecting the Twitter data. R is a programming language and environment for statistical computing and graphics representation [30]. R was chosen for its simplicity when it comes to data collection, data storing and statistical computations.

Rtweet Rtweet is a package that facilitates the connection to the Twitter search API. It has predefined functions which helps the user to search for specific Twitter data. This package has been used together with R, in order to make the data gathering easier.

3.2 Data Preprocessing

The preprocessing stage is critical in order to obtain a high accuracy in the sentiment analysis [10]. Before the data is cleaned there are a lot of spam, special characters and other unnecessary data which not are required or can be classified in the analysis. Tweets were classified as spam if a user tweeted the same content several times. These tweets were deleted during this stage. Then blank spaces at the beginning and the end, usernames, punctuations, links, tabs and tweets made in other languages than english were deleted.

3.3 Sentiment Analysis

After the preprocessing stage, the data is stored as an R data frame since the following work also was done in R. Sentimentr was selected to perform the sentiment analysis. The package is well used to calculate text polarity on a row-level. Sentimentr also provides the opportunity to take valence shifters in consideration, e.g negators, adversative conjuctions, amplifiers and de-amplifiers. The words of each sentence is compared to a polarity dictionary and then tagged with a number between -1 and +1 where -1 is negative and +1 is positive. Sentimentr have weights on valence shifters and the standard weights

provided by the creator was chosen. The final result is the divided by the number of words in order to get a value which is not dependent on the length of the sentence.

3.4 Data Aggregation

When the sentiment analysis was completed, the daily sentiment was calculated for each stock during the period. This was done by calculating the mean value of all sentiments from all tweets during each day. One day is defined as one minute after the closing of the stock exchange until the next day's closing. The sentiment values was inserted into a matrix with two columns where column one contained the stock's daily sentiment, and where column two contained the close price of the stock on the corresponding day.

3.5 Correlation Calculation

The correlation was calculated from the values in the matrix containing the sentiment and the stock prices. The correlation r was calculated using the Pearson correlation coefficient:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2(y_i - \bar{y})^2}} \tag{1}$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i \tag{2}$$

and n is the number of days, x_i and y_i are the values from the sentiment and the stock prices.

The correlation was calculated by using R's built-in function *cor*.

3.5.1 Accuracy and Precision

Once the correlation has been calculated, it is analyzed how stable and credible the correlation value is. This consists of three parts: statistical significance, shuffling and stability.

Statistical Significance The p-value is the probability that the current result were found if the correlation coefficient were in fact zero. If this probability is lower than the usual 5% ($P < 0.05$) the correlation can be seen as statistically significant [29]. The p-value was calculated for each correlation coefficient in this report. This was done in R.

Shuffling test The data points are randomly shuffled in several tests where the correlation are computed for every test. Then the mean value of these tests was calculated.

Stability Once the correlation value was calculated, the stability in the correlation was analyzed. This was done by examining how much the correlation value was affected if one or more days were removed in the calculation. A stable correlation value should not be affected to a greater extent if certain values are removed.

4 Results

This chapter will present the results that has been found in the collected stock and twitter data, with the statistical methods applied to them. The results are illustrated both in graphical and tabular form together with explanations. A selection of stocks from different industries will be presented in detail, but results from all companies can be found in the summary.

Sample tweets are also shown for each stock. These tweets have been randomized from the saved R data frames to give a picture of the data used to calculate the correlation, together with the sentiment score. Links and usernames have been deleted from these tweets.

4.1 Apple

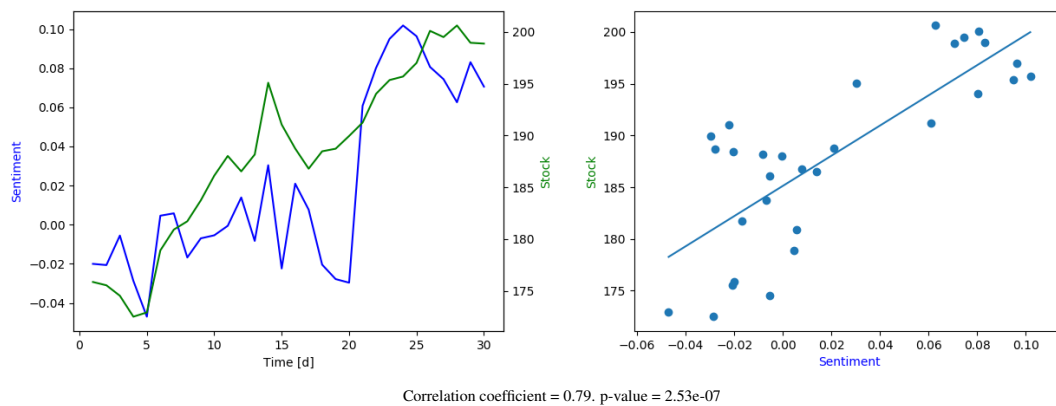Correlation coefficient = 0.79. p-value = 2.53e-07

Figure 4.1: Illustration of the correlation for $AAPL.

Figure 4.1 shows that it exists a correlation between the Twitter sentiment and the Apple stock. The correlation can also be considered as stable. The value is close to the original value if the correlation is calculated with different selected sub intervals.

$AAPL	
Tweet	**Score**
Somebody credible should spread negative news on $AAPL so that I can buy more on the dip.	-0.06
LOL. But loves $aapl now.	1.00
LOL! Stay sharp friend. See my last $AAPL post for the Weekly chart	0,50
No real value investor would use $AAPL anything	0
$aapl aapl looking great...amazing option plan by Sandro!	0,47

Table 4.1: Sample of the collected tweets containing the text segment $AAPL.

4.2 AIG

Figure 4.2 shows that it does not exist a correlation between the Twitter sentiment of AIG and the stock price.
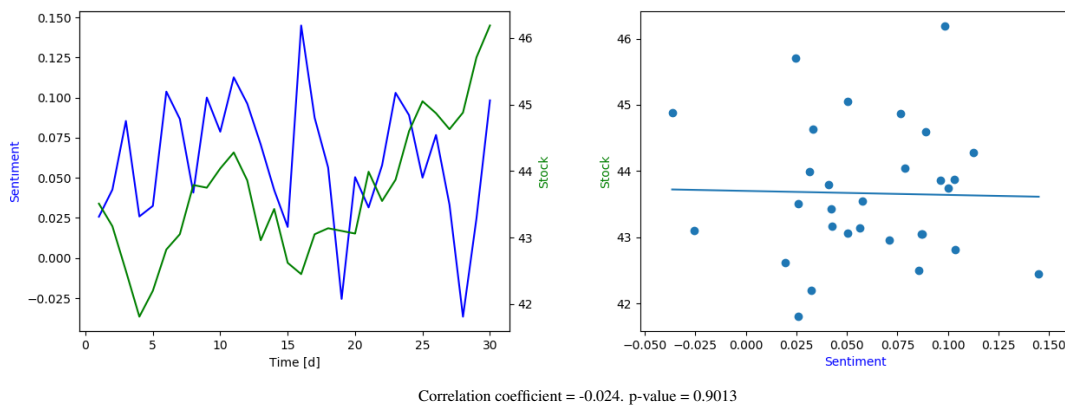

Correlation coefficient = -0.024. p-value = 0.9013

Figure 4.2: Illustration of the correlation for $AIG.

$AIG	
Tweet	**Score**
Eyes on $AIG	0
$AIG with an insidious move above 45...will it hold?	-0,17
$AIG Book Value is $65.04 - Ready to run.	0.30
Analysts See $1.08 EPS for American International Group, Inc.	0
$AIG Bullish above $42.50	-0,46

Table 4.2: Sample of the collected tweets containing the text segment $AIG.

Table 4.2 shows a sample of the tweets that have been used for this stock. From this, one can see that the tweets are relevant to the stock. However, there were relatively few tweets containing the AIG ticker, only 909 tweets, which is only around 25 tweets per day, compared to the Apple ticker that had 55.691 tweets during the same period. This may be a reason for the low correlation value since every tweet had a high impact on the average sentiment score for each day.

4.3 Boeing

4.3.1 Boeing

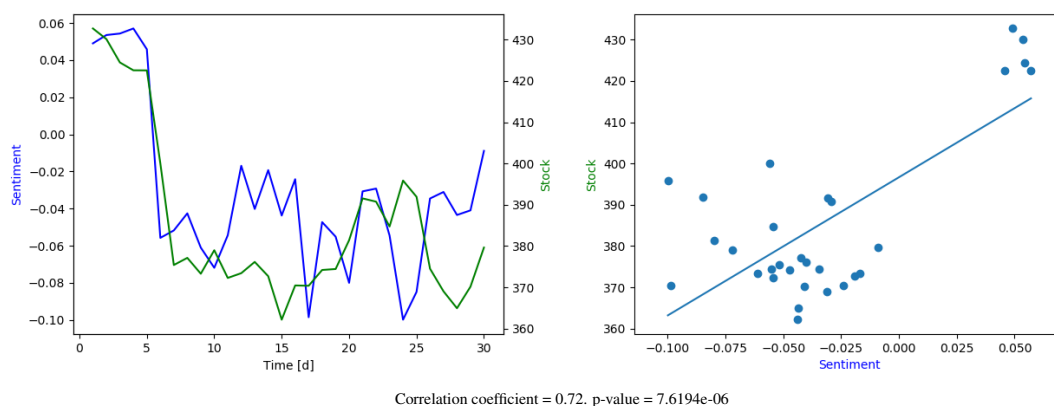Correlation coefficient = 0.72. p-value = 7.6194e-06

Figure 4.3: Illustration of the correlation for Boeing.

Figure 4.3 shows the sentiment of the tweets that contain the key word "Boeing". The figure shows that there exists a correlation between the Twitter sentiment and the stock price. The dip at the beginning of the time interval shows when a Boeing aircraft crashed

in Ethiopia on Sunday March 10, 2019 [11]. This caused a negative outcome of the stock and the sentiment. Although the correlation value can be considered good enough to be able to claim that there is a correlation, the correlation value is not stable. If the correlation is calculated with only the values after the crash, the correlation value will decrease significantly.

Boeing	
Tweet	**Score**
Update Max8: Dennis Tajer, a spokesman for the American Airlines pilot union and a 737 pilot, said the training and other information had restored pilots confidence in @Boeing since the Indonesia crash.	0.15
I used to be boeing's fan but i don't like at how they are handling the situation on the crashed of its 2 planes 737 max 8, i guess i prefer airbus now	-0.30
Fix this shit NOW @boeing.	-0,34
Boeing pauses 737 Max deliveries as probe into Ethiopian airlines black box gets underway.	-0,07
BBC News – American Airlines extends Boeing 737 Max flight cancellations.	0,13

Table 4.3: Sample of the collected tweets containing the text segment Boeing.

4.3.2 $BA

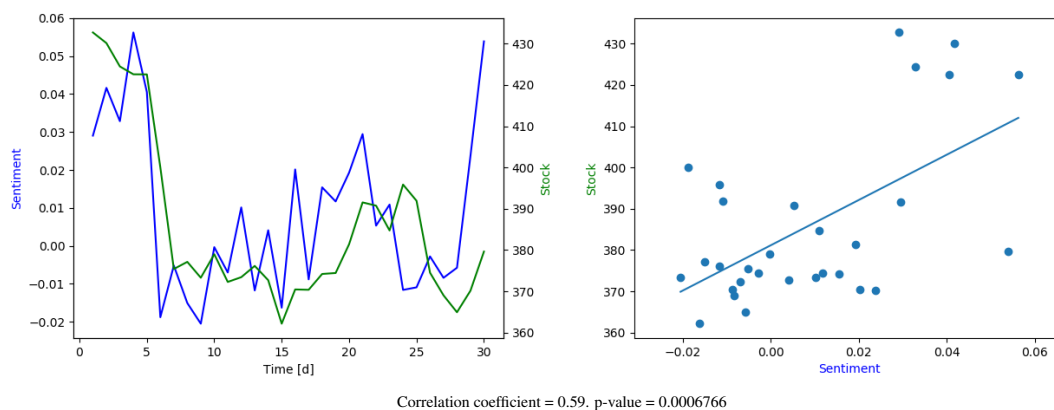Correlation coefficient = 0.59. p-value = 0.0006766

Figure 4.4: Illustration of the correlation for $BA.

Figure 4.4 shows the correlation when calculating the sentiment value of tweets containing the ticker "$BA". The correlation value of the ticker is not as good as when using the key word "Boeing", but it is more stable when testing the same sample days.

$BA	
Tweet	**Score**
Boeing $BA has been by far the best stock in the Dow YTD, +31%. Monday's trading will be a critical test of bullish sentiment.	-0,19
$BA: Indonesia airline seeks to cancel 737 MAX 8 jet order.	-0,25
Boeing $BA could get very ugly.	-0,55
$BA no one asking why airlines were too cheap to buy the extra safety features! Theyre at fault too for putting customers life at risk!	-0,11
The best deal will be $BA at open	0,27

Table 4.4: Sample of the collected tweets containing the text segment $BA.

The sample tweets from Boeing are more relevant to the company in general, while the $BA sample tweets are more relevant to the stock in particular. The tweets related to the stock gives a more stable correlation than those related to the company as a whole.

4.4   Facebook

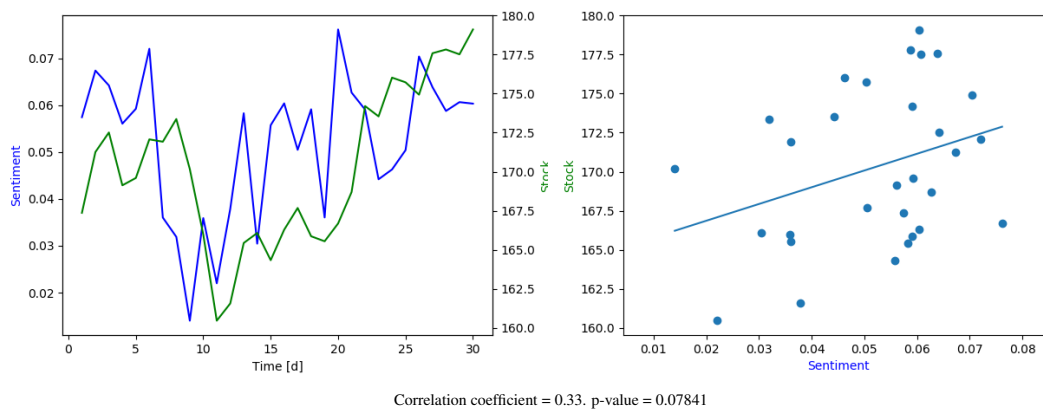Correlation coefficient = 0.33. p-value = 0.07841

Figure 4.5: Illustration of the correlation for $FB.

Figure 4.5 shows that there exists a small positive correlation between the Twitter sentiment and the stock.

$FB	
Tweet	**Score**
$FB Im shorting - zucks and company are unloading huge here they know	0
Investors Buy Facebook $FB on Weakness After Insider Selling	0,33
$FB bullish over 180. Probably headed to yellow line above	-0,27
$GOOGL bull/bear line is that 1200 level. Not out of the woods yet. Could get slaughtered with a head and shoulder smack down if it doesnt cross. Weak day today is a red flag. Maybe it has $FB type problems directly ahead? No matter. 1200 level speaks for itself.	-0,71
$FB gonna pay huge tomorrow. These bullish daily candles are always followed by another huge big candle.	-0,28

Table 4.5: Sample of the collected tweets containing the text segment $FB.

The sample shows that the first tweet is strongly negative, but that it gets score 0 since it contains finance slang word that the sentiment model is unable to interpret correctly.

4.5 Home Depot

4.5.1 $HD

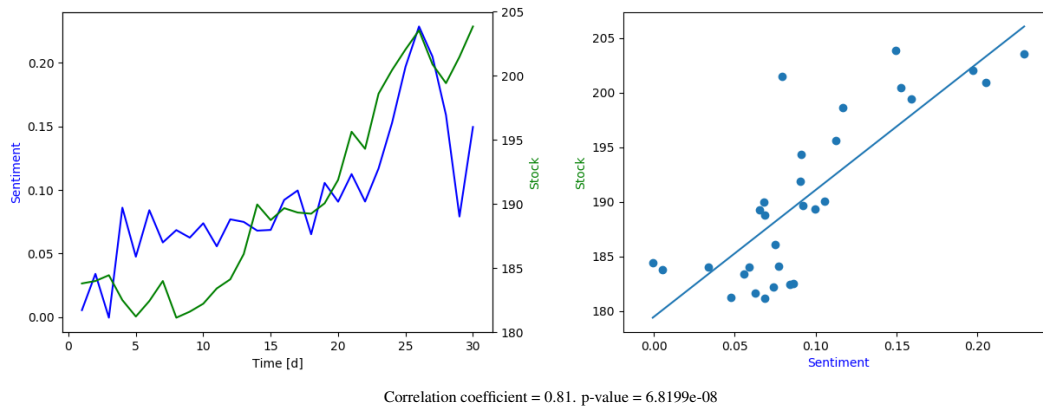Correlation coefficient = 0.81. p-value = 6.8199e-08

Figure 4.6: Illustration of the correlation for $HD.

Figure 4.6 shows that it exists a correlation between the Twitter sentiment and the Home Depot stock.

$HD	
Tweet	**Score**
2019-04-12 Short sale volume (not short interest) for $QQQ is 58%. $HD 51% $X 49% $FEYE 46% $VTL 51%	-0,14
Home Depot $HD Holder Kelly Lawrence W & Associates Upped Holding; As Gildan Activewear $GIL Market Value Rose, Causeway Capital Management Has Cut Its Position by $4.89 Million	0,03
$HD Home Depot Inc. Option Order Flow Sentiment is 54.1% Bullish.	-0,13
As Home Depot $HD Stock Rose, Holder Wright Investors Service Trimmed Holding by $1.16 Million; Granite Constr $GVA Holder Fiduciary Financial Services Of The Southwest Lowered Its Stake.	0
$HD stay away for now. study All 3 circles red. 1. Underperforming 2. RSI below 50 3. Price has been rejected near short term ema.	-0,17

Table 4.6: Sample of the collected tweets containing the text segment $HD.

4.5.2 Home Depot

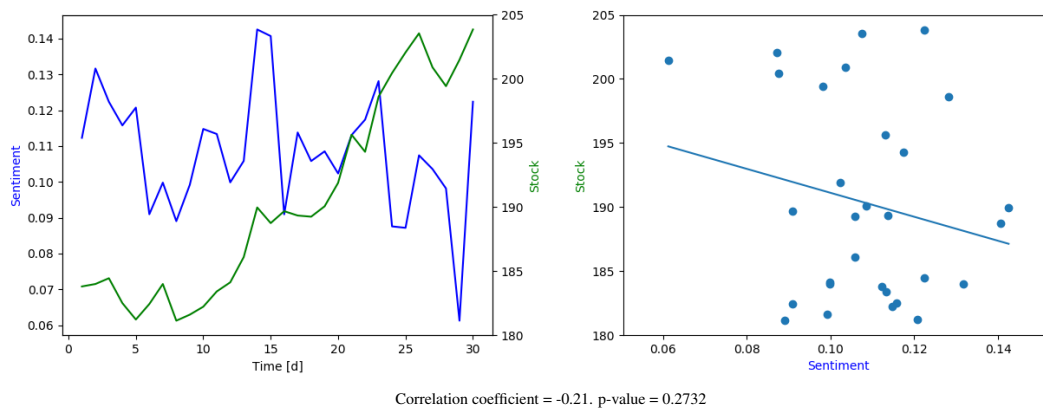

Correlation coefficient = -0.21. p-value = 0.2732

Figure 4.7: Illustration of the correlation for Home Depot.

Figure 4.7 shows that there is a weak, negative correlation between the Twitter sentiment for the key word "Home Depot" and its stock price.

Home Depot	
Tweet	**Score**
`I basically live at Home Depot.`	0
`Home Depot appliance return policy sucks`	−0,2
`getting home depot for lunch`	0
`Some dude at Home Depot got his truck broken into, they stole his iPad and his Lunch!`	−0,37
`Enter To Win A $50 Home Depot GiftCard (10 Winners) Sweeps Ends 4-17 SH`	0,31

Table 4.7: Sample of the collected tweets containing the text segment Home Depot.

The tweets with the key word "Home Depot" shows completely irrelevant data that has nothing to do with the stock, which resulting in a low correlation value, as shown in figure 4.7.

4.6   Netflix

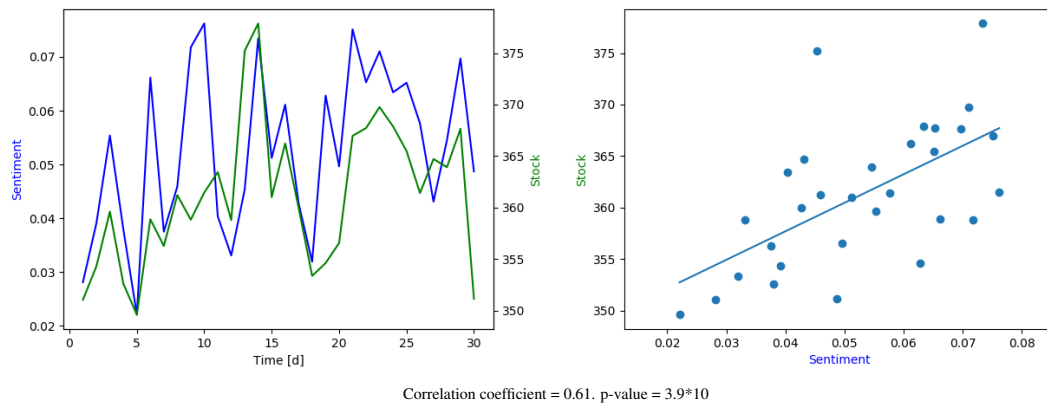Correlation coefficient = 0.61. p-value = 3.9*10

Figure 4.8: Illustration of the correlation for $NFLX.

Figure 4.8 shows that it exists a correlation between the Twitter sentiment and the Netflix stock.

$NFLX	
Tweet	**Score**
$NFLX Netflix, one of the great winners of all time, looks risky ahead of earnings, warns technician	0,33
$NFLX Can Netflix keep its subscribers? from paching by $DIS	0
$NFLX nice gap up this morning. Worth watching.	0,44
$NFLX $GOOGL: Netflix Will Beat Expectations for Subscriber Growth Again, Analyst Predicts.	0,22
Other holdings mixed to positive. Long $CHK and long $SB were strong. Long $GLD gave some back. Short $NFLX moved against me slightly. Best ideas tweeted live. #fintwit	0,26

Table 4.8: Sample of the collected tweets containing the text segment $NFLX.

The sample tweets for "$NFLX" are relevant to the stock. Although, some financial slang is used in some tweets which might affect the score on those particular tweets.

4.7 Summary of the Result

The results shown in table 4.9 indicates that there are some cases with a correlation between the stock performance and the Twitter sentiment. The conducted shuffling tests showed that the correlation coefficients for the respective stocks are significant and the p-values are within the expected confidence intervals. The mean correlation of the shuffled data sets shows no correlation which further proves that the correlation is not an random event.

Key word	Number of collected tweets	Correlation	p-value	Shuffled correlation
$HD	13.261	0,81	$6,8*10^{-8}$	0,004
$AAPL	55.691	0,79	$2,5*10^{-7}$	0,034
Boeing	177.287	0,72	$7,6*10^{-6}$	-0,015
$NFLX	41.484	0,61	$3,9*10^{-4}$	-0,011
$BA	30.786	0,59	$6,8*10^{-4}$	0,015
$FB	37.183	0,33	$7,8*10^{-2}$	0,051
$AIG	909	-0,024	0,90	0,13
Home Depot	51.863	-0,21	0,27	0,08

Table 4.9: Correlation and p-values for each key word, sorted by correlation value.

Calculating the average correlation coefficient per sector gave the results illustrated in table 4.10. The results shows no clear difference between the sectors except from the food sector.

Sector	Average correlation
Technology	0,175
Industrial	0,140
Manufacturing	0,133
Finance	0,125
Healthcare	0,075
Food	-0,194

Table 4.10: Correlation coefficients by sector, sorted by correlation.

The distribution of correlation coefficients show that the majority of companies have a coefficient in the range 0,16 - 0,22, which is shown in figure 4.10.
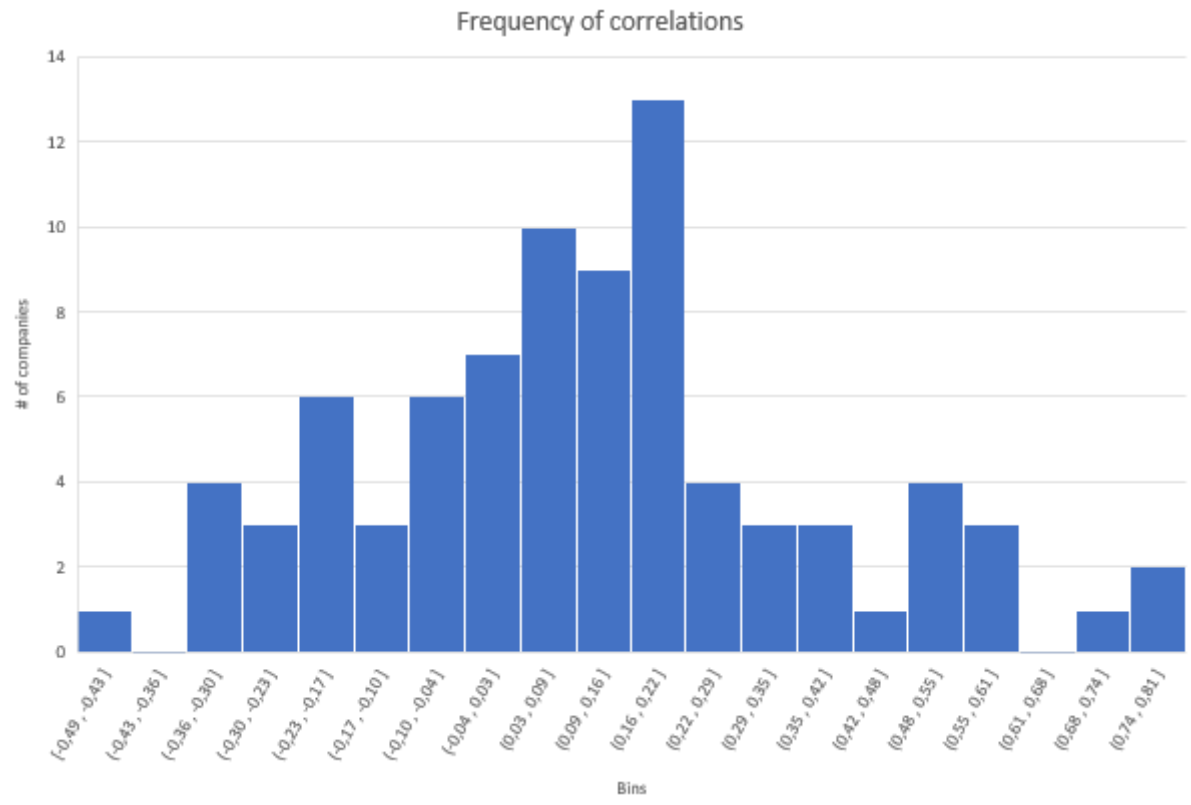

Figure 4.10: Histogram showing the frequency distribution of correlation coefficients from all companies

5   Discussion

This chapter will include an analysis of results, reliability, limitations and ethical issues.

5.1   Analysis of Results

The found results shows that there are some cases with a correlation between the stock price and the Twitter sentiment, however it is dependent on several different factors. In order to find correlation between the price and the sentiment, the quality of data is critical. For example, when comparing the results of the key word "Home Depot" and its stock ticker "$HD", they illustrate how quality of data affects the correlation. The sample tweets from key word "Home Depot" has no relevancy to the stock or the company's performance. The sample instead shows tweets about events surrounding Home Depot and other miscellaneous events. On the contrary, the "$HD" sample tweets has high relevancy to the stock which results in a high correlation coefficient. The same patterns can be seen when looking on key word "$NFLX" where all sample tweets are relevant to the stock. On the other hand, the quality of AIG's sample tweets are high as five out of five tweets are relevant to the stock, while the correlation coefficient is -0.0024, which is no correlation at all. Looking more closely into the data, one can find that there was a total of 909 tweets gathered on the key word "$AIG" and as a comparison, "$AAPL" had 55.691 tweets during the same period. The weight of every tweet gets higher when the amount of data points are fewer. Therefore, in the "$AIG" case, every single tweet affects the result to a greater extent which could make one question the reliability of this particular data.

During the period of the data gathering, a Boeing airplane crashed in Ethiopia, which significantly affected the stock price and the sentiment. Looking more closely into the data surrounding the crash, one can see big differences in the correlation coefficient. Measuring the first 15 days of the 30 day period, the correlation coefficient is 0.89 for the key word "Boeing" and 0.83 for key word "$BA". Measuring the last 15 days of the period the correlation coefficient is -0.29 for key word "Boeing" and 0.13 for the key word "$BA" showing unstable values over time. This could imply that a significant event increases the correlation between the sentiment and the stock price. A plausible explanation for this could be that during such events, the weight of traditional and rational valuation

techniques decreases, while the public opinion more reflects the stock price.

Calculating the correlation coefficient per industry shows no clear difference between industries except from the food sector which had an negative average and all evaluated companies had a negative correlation coefficient. The companies in the food sector consisted of only four companies compared to 22 in the technology sector, in order to draw any further conclusions one would need to extend this data set.

5.2 Reliability of Sentiment Score

The results indicate some unreliability in the sentiment score, which could be derived mainly from the quality of data. Some key words retrieved tweets which was discussing other stocks and irrelevant information. This is a challenge in NLP, since a more thorough data cleaning could result in significantly fewer data points and also the risk of cleaning relevant data.

Even-though the NLP dictionary consisted of hundred-thousands of words, there is a vast amount of industry specific slang used in the tweets which the sentiment fails to score correctly. In table 4.5 showing the sample tweets of Facebook, one tweet says "Im shorting - zucks and company are unloading huge here they know", which basically means that the author believes that the stock price is going down since Mark Zuckerberg and others insiders sell their shares. Shorting and unloading have their industry specific meanings which is not taken into account for in the dictionary and even-though this is a very negative tweet, it gets a neutral score of zero.

5.3 Limitations

The primary factors that have limited the thesis consists of *time, data* and knowledge in the field of *Natural Language Processing*.

Our work was restricted by a limitation of Twitter data, since the Twitter API only can search for tweets up to seven days back in time and not more than 18.000 tweets every fifteen minutes. In order to get full access to the Twitter API, the user has to pay 2499$/month which limits this service mainly to corporations. This has led to a limited amount of data during a short period. A longer period of time is necessary in order to to get a better result.

The majority of the collected tweets contained slang words and emoji symbols, which the sentiment analysis could not handle in a complete manner. An improvement of this could lead to significantly better results.

Our own knowledge within NLP limited us since we could not change the algorithms or dictionaries to suit our field of study. If our knowledge would have been better we could have optimized the dictionary to the financial sector which would have increased the reliability of the sentiment scores.

5.4   Ethical Issues

Even if this thesis does not raise any ethical concerns, one might be cautious and discuss what kind of behaviour it might lead to in the future. As mentioned in the introduction the amount of hedge funds using alternative data sources has increased over the last few years [13] and the search for a competitive edge on the stock market might lead to unethical behaviours in terms of privacy.

6 Conclusion

One cannot say that there is a clear correlation between the Twitter sentiment and the stock price. In order to get more reliable results much more data would be needed and also a better method to clean up the data. There are some evidence suggesting that if the quality of the data is good then there is a correlation between the Twitter sentiment and the stock price. But in most cases, there are other factors driving the stock price. Also, the results suggests that the importance of sentiment is fluctuating and could increase if a significant event has occurred which might affect the stability of the correlation. There are no or little evidence showing that the correlation of Twitter sentiment and stock price has any relation to a specific industry or product type.

Regarding the certainty of our results, the correlation seems to be significant, but the found results does not prove causality between the two. An investor should not solely base their analysis on these results but it could be used as an extra layer in their analysis improving accuracy or to monitor company sentiments in long term, considering the high correlation on the clean data sets.

6.1 Future Research

Future challenges in the field based on this thesis are primarily to estimate the informal language on social media that primarily consists of slang and emoji symbols, which sometimes can be critical when determining the sentiment value for a sentence.

References

[1] *About Apple Inc.* https://www.reuters.com/finance/stocks/company-profile/AAPL.O. 2019.

[2] *About Boeing.* https://www.boeing.com/company/general-info/. 2019.

[3] *About Facebook.* https://www.reuters.com/finance/stocks/company-profile/FB.OQ. 2019.

[4] *About Home Depot.* https://www.reuters.com/finance/stocks/company-profile/HD.N. 2019.

[5] *About Netflix.* https://www.reuters.com/finance/stocks/company-profile/NFLX.OQ. 2019.

[6] *AIG - About us.* https://www.aig.com/about-us. 2019.

[7] Blom, Gunnar et al. *Sannolikhetsteori och statistikteori med tillmpningar.* Studentlitteratur AB, 2017. ISBN: 9789144123561.

[8] Buffett, Warren. "The Superinvestors Graham-and-Doddsville". In: (1984).

[9] Damodaran, Aswath. *Session 7: Market Efficiency - Laying the Groundwork.* 2014.

[10] Emma Haddia Xiaohui Liua, Yong Shib. *The Role of Text Pre-processing in Sentiment Analysis.* https://www.sciencedirect.com/science/article/pii/S1877050913001385. 2013.

[11] *Ethiopian 737 Max Initial Crash Report Set to Be Released Thursday.* https://www.bloomberg.com/news/articles/2019-04-03/ethiopian-737-preliminary-crash-report-to-be-released-thursday. 2019.

[12] GEORGIOS PALTOGLOU, MIKE THELWALL. *Twitter, MySpace, Digg: Unsupervised Sentiment Analysis in Social Media.* 2012.

[13] Hwang, Hailiang Chen Prabuddha De Yu (Jeffrey) Hu Byoung-Hyoun. *Article Navigation Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media.* 2014.

[14] Investopedia. *Forces That Move Stock Prices.* https://www.investopedia.com/articles/basics/04/100804.asp. 2019.

[15] Investopedia. *Stock Market.* https://www.investopedia.com/terms/s/stockmarket.asp. 2019.

[16] Johan Bollen Huina Mao, Xiao-Jun Zeng. *Twitter mood predicts the stock market*. `https://arxiv.org/pdf/1010.3003.pdf`. 2010.

[17] Kenton, Will. "Random Walk". In: (2018).

[18] Lee, Mike. *How global hedge fund managers can embrace innovation*. Ernest Young, 2017.

[19] Mao, Yuexin et al. "Correlating SP 500 stocks with Twitter data". In: Aug. 2012, pp. 69–72. DOI: `10.1145/2392622.2392634`.

[20] Nofsinger, John R. "Social Mood and Financial Economics". In: *Journal of Behavioral Finance* 6.3 (2005), pp. 144–160. DOI: `10.1207/s15427579jpfm0603 \ _4`. eprint: `https : / / doi . org / 10 . 1207 / s15427579jpfm0603_4`.

[21] Pagolu, Venkata Sasank et al. *Sentiment Analysis of Twitter Data for Predicting Stock Market Movements*. `https://arxiv.org/pdf/1610.09225.pdf`. 2016.

[22] Shuyuan Deng Zhijian James Huang, Atish P. Sinha and Zhao, Huimin. *The Interaction between Microblog Sentiment and Stock Return: An Empirical Examination*. 2018.

[23] Smith, Aaron and Andersson, Monica. "Social Media Use in 2018". In: (2018).

[24] Thelwall, M. *MySpace Comments. Online Information Review*. 2009.

[25] Torsten Schmidt, Simeon Vosen. *Forecasting Private Consumption: Survey-based Indicators vs. Google Trends*. `https://pdfs.semanticscholar.org/66f2/b2cfe56fc2414d3036cb6e28098bfa55216f.pdf`. 2009.

[26] Tumasjan, Andranik. *Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment*. 2010.

[27] Twitter. *Twitter Announces Fourth Quarter and Fiscal Year 2018 Results*. `https://s22.q4cdn.com/826641620/files/doc_financials/2018/q4/Q4-2018-Earnings-Pres-Release.pdf`. 2019.

[28] Twitter. *Twitter Search API*. `https://developer.twitter.com/en/docs/tweets/search/overview`. 2019.

[29] Wasserstein, Ronald L. and Lazar, Nicole A. *The ASA's Statement on p-Values: Context, Process, and Purpose*. `https://www.tandfonline.com/doi/full/10.1080/00031305.2016.1154108`. 2016.

[30] *What is R?* `https://www.r-project.org/about.html`. 2019.

Appendices

A   Twitter Keywords

The following key words were used to collect the Twitter data. The search terms for each stock is the company name, the stock ticker including a dollar sign, and in some cases the stock ticker without a dollar sign.

A.1   AIG

- American International Group
- $AIG
- AIG

A.2   Apple

- Apple
- $AAPL
- AAPL

A.3   Boeing

- Boeing
- $BA

A.4   Facebook

- Facebook
- $FB

A.5   Home Depot

- Home Depot

- $HD

A.6   Netflix

- Netflix

- $NFLX

- NFLX