Jens-Otto Forbergskog
Christer Ryland Blom

Master Thesis

Twitter and Stock Returns

Date of submission:
02.09.2013

Supervisor:
Prof. Øyvind Norli

BI Norwegian Business School

Oslo

Master of Science in Business and Economics, Major in Finance

GRA 19003 - Master Thesis

Acknowledgements

Abstract

In this thesis, we investigate whether the sentiment of tweets mentioning stock tickers can be used to predict stock performance. In particular we test for leading and lagged relationships between the percentage of positive and/or negative tweets and the returns of the S&P 500 index. We obtain a longitudinal data set of all tweets mentioning stock tickers over a four-month period amounting to 2,599,277 tweets distributed over 84 trading days. We use daily measures for positive and negative sentiment to generate our explanatory variables. Our results indicate that an increase in the percentage of positive tweets predicts increased stock performance the following day whereas an increase in the percentage of emotional tweets predicts a reduction in stock returns after two and three days. An increase in the percentage of negative tweets may predict a reduction in stock returns.

Table of Contents

List of Figures

List of Tables

Introduction

The use of social media, such as Twitter, by financial market participants is a recent phenomenon and is therefore poorly covered by academic research. We believe that social media will play an important role in the lives of investors, traders, and analysts in the future. Therefore, we feel it is an area worthy of academic study.

This thesis examines whether there is any valuable information concerning stocks shared on Twitter. Specifically, we are testing for relationships between the sentiment contained in stock specific tweets and stock returns. Our findings suggest that an increase in the percentage of positive stock specific tweets predicts increased stock returns the following day, while an increased level of emotionality can predict reduced stock returns two to three days in advance. Increased stock performance as well as an increase in the share of negative tweets can predict an increase in the share of positive tweets. In the longer term, we find that a one standard deviation increase in the percentage of negative tweets leads to a 0.44 percentage point cumulative reduction in stock returns after seven days.

The use of big data analysis (i.e. analysis of huge amounts of unstructured raw data) by investors has already begun. Indeed,

> some of the largest quant hedge funds, the likes of Renaissance Technologies, D.E. Shaw and others are said to be spending millions (if not billions) on building tools for analyzing unstructured data found on Twitter and Facebook. Big data companies like Thomson Reuters and Dow Jones are offering products and entire business units around interpreting sentiment analysis to produce trading signals. (Schmerken 2012)

It is clearly big money in analyzing big data. Some of the most enthusiastic advocates of social media are of the opinion that it can be "construed as a form of collective wisdom" (Asur and Huberman 2010), and being able to tap into this collective wisdom, should make for superior investment decisions.

Automated trading based on algorithms analyzing real time market data from the stock exchanges is now quite common and well covered by research and media. Most of these algorithms are however exclusively analyzing data provided by the stock exchanges themselves. Imagine instead a sophisticated algorithm able to tap into the entire pool of information known to man, including news feeds and real

time discourse in social media, then based on this make instantaneous, unbiased and rational investment decisions. We believe this could become the next generation of algorithm trading which may replace not only traders, but also brokers and analysts. If this becomes a reality, computers will most definitively have an even more prominent role in the stock markets than they have today.

As social media grows in popularity, an increasing share of valuable information is shared there. In addition to the increase in the amount of information, the credibility of Twitter has gradually increased as more prominent organizations have accepted and adopted it. News organizations were among the first to embrace Twitter as a channel to spread news. Now, it is hard to find any newspaper or journalist without a Twitter account.

Also, regulators have taken note of the trend towards social media as a source of market information. On the 2nd of April 2013, the Securities and Exchange Commission (SEC) approved social media as an official communication channel for stock sensitive information (SEC 2013b). Two days later, on the 4th of April 2013, Bloomberg announced they would integrate live Twitter feeds into their terminals, further strengthening Twitter's position (Bloomberg 2013b).

Increased usage of social media as a source for investment decisions in combination with the large and ever increasing amount of available data have led to debates on the level of information content of these public media. Does tangible economic information exist in these data sources, or is it all just noise? Until recently it was not possible to analyze big data due to computational, storage and bandwidth constraints. The human brain has so far remained superior to computers in understanding the broader picture, but rapid technological development infers that machines will eventually prevail in such analysis.

We continue by reviewing current academic literature on behavioral finance, market efficiency, big data, social media, and sentiment analysis. We then introduce our hypotheses, data set, and methodology before presenting our results and conclusions.

Literature Review

Twitter is an online social media service used by millions of individuals and organizations worldwide to exchange short messages of up to 140 characters. It has rapidly evolved over the past few years to become a complete ecosystem and a powerful tool in several areas such as news, politics, health, and in our case, finance.

The growth of Twitter since its conception in 2006 has been extraordinary. There are now officially over 200 million active Twitter users (Twitter 2013), and tweet volume has grown significantly the past years, from 230 million daily tweets at the end of 2011 (Lane 2012) to more than 400 million tweets per day in November 2012 (PeopleBrowsr 2012).

Twitter was originally intended as a rapid message service for emergency personnel (Bloomberg 2013a). The idea was that first responders could tweet their status and location to help others decide their most appropriate action. One of the key elements separating Twitter from most earlier messaging services is that there is no specified recipient; all subscribers receive the information in real time. Each user decides whom to follow and thus receives a unique stream tailored to his or her interests.

The Anatomy of Twitter

The core of Twitter is called the *firehose* which is the constant stream of all tweets. Each *tweet* contains a text message of up to 140 characters, with additional embedded metadata such as author, time and date, location, and language. Table 1 below describes the main Twitter terms and concepts.

Hashtags, *at*, and *cashtags* are used as text modifiers to create structure. When a word is assigned any of these modifiers by putting it directly in front of it (e.g. #earnings), users can click on them to find related tweets. Clicking on a username shows the user's profile with all previous tweets. Clicking on a cashtag or hashtag shows recent tweets mentioning the tag.

Table 1 - Twitter Terms and Concepts

Twitter	The brand and company
Tweet	An up to 140 character long text message
Firehose	Constant stream of all tweets in real time
Hashtag (#)	Identifies a topic (e.g. #earnings)
At (@)	Identifies a username (e.g. @CNBC)
Cashtag ($)	Identifies a stock ticker (e.g. $GS)
Followers	Users who subscribe to tweets sent by a user
Mentions	Number of times others mention a specific username
Retweet	When another user relays your tweet to their followers

Even though tweets in principle are public, direct access to the firehose is restricted and has recently been tightened. We see two main reasons for this. First, the enormous amount of data makes it incredibly difficult and expensive to serve this stream to everyone. Second, Twitter likely realized that selling access to the full stream had the potential to become one of their main sources of income. They have recently started to cancel early firehose access contracts signed before they knew how popular their service would become. This move was unpopular, and one early partner, PeopleBrowsr, even launched a lawsuit to retain their access. They won a restraining order in November 2012 forcing Twitter to continue providing firehose access (PeopleBrowsr 2012). Microsoft seems to have paid the USD 30 Million yearly fee Twitter reportedly demanded to provide firehose access for their search engine Bing. Negotiations with Google, however, failed and they are therefore no longer offering real time Twitter search (Gannes 2011). There are some select distributors such as Gnip and DataSift who are licensed to relay their firehose access to third parties (Lane 2012). These distributors also keep complete records of all tweets which can be accessed for historical data analysis. Also, companies such as Dataminr pay for full firehose access directly and sell real time analytics to financial firms and governments.

Due to its increasing popularity and credibility, Wall Street has shown increased interest in Twitter. On the 4th of April 2013, Bloomberg announced that they would incorporate Twitter streams into their terminals. Bloomberg provides additional functionality, most notably filter capabilities enabling users to filter by companies, industries, markets, and people. They also provide lists over trending companies on "Bloomberg Social Velocity" (Bloomberg 2013b). An example of the user interface is shown in Figure 1 below.

Figure 1 - Bloomberg's Twitter Integration

Research on Twitter

Several different streams of research on Twitter can be identified. One stream focuses on understanding its usage and community structure. Another focuses on the influence Twitter users have, for example by changing the outcome of an election, exposing unethical behavior by companies, uncovering scandals, and affecting product sales. Our research belongs to a third stream which focuses on Twitter's prediction power and potential application to other areas.

The usage and community structure of Twitter have evolved over time. The early adopters were, as usual in the case of new technology, computer geeks. After reaching a critical mass, reporters and politicians began to see the value of using Twitter to spread news and political messages. Some argue that the usage of Twitter and other social media was one of the reasons Obama won the US presidential election in 2008 (Fraser and Dutta 2008).

The last two streams are interrelated. It is not always obvious whether tweets are affecting external factors, or are being affected by them. Although methodologically hard to prove, intuitively it makes sense that tweets can affect the profitability of companies. Twitter is one of many sources where consumers seek information before potential purchases. If they discover that other consumers have had significant negative experience with the company, they may refrain from purchasing the company's product, lowering sales. On the other hand, prior

negative performance may affect the sentiment of tweets, giving origin to the feedback effect often observed.

What Makes Twitter Different from Other Media?

News wires function in similar ways to Twitter, with the notable difference that all posts originate from news organizations. The providers (such as Reuters and Bloomberg) broadcast their streams directly to subscribers, as well as to several third party services who relay this information to end users. In addition, most provide the news stream on their websites. News streams have been around for decades, thus there has been ample academic research conducted on their role in the financial markets. They remain one of the most important sources of information for market participants, providing timely and reliable news.

Discussion forums can be seen as the predecessors of modern social media. While discussion forums remain in use today, technological and infrastructural advances have to a certain degree attracted users toward more modern conceptions of social media such as Twitter, whose design and structure facilitates and speeds up information exchange. Whereas discussion forums require users to actively pursue topics, modern social media instead pushes information to the user based on certain criteria. Stock message boards are often characterized as places where individuals can seek, gather and discuss information and opinions on various stocks, and are usually available to the public.

Following significant increases in the usage of stock message boards as discussion forums towards the end of the 1990s, Wysocki (1998) investigated the relationship between message posting volume and firm characteristics and stock market activity. He found that message posting volume predicted future trading volume and stock returns. Antweiler and Frank (2004) found that messages generated on stock message boards "help predict market volatility" while also having a statistically significant, albeit economically small, effect on stock returns.

Compared to the forums that predate modern social media for user-generated content online, Twitter has powerful self-moderating aspects that users appreciate:

> By letting each individual user decide whom to follow, the content is moderated automatically: 'Underperformers will be ignored, and rightly so—trading is a zero-sum game and bad advice is a waste of time and money. That's precisely what validates apps like StockTwits'. (Zeledon 2009)

This means that users who provide useful information become more visible, while spammers are kept in the shadows, which should provide for a higher signal to noise ratio. In addition, trusted sources such as major news agencies, governments and companies have verified accounts so users can trust that the information in tweets from such sources is backed by more than just a nickname.

As Twitter has grown in popularity, its user base is now not only counting individual users, but has also become increasingly more important as an arena for organizations, businesses and public services. Kaplan and Haenlein (2010) discuss reasons for the increased success of micro blogs such as Twitter, noting that their successes come as a result of their unique communicational characteristics, resulting in; "the creation of ambient awareness; a unique form of push-push-pull communication; and the ability to serve as a platform for virtual exhibitionism and voyeurism". Whinston and Rui (2010) argue that "the unique innovation of social media is recognizing and connecting people's need for information and attention" and as such, its design should facilitate such a connection. Whinston and Rui also find that whether or not a user becomes a content producer or consumer depends on the relationship between their reservation wages for becoming either a producer or a consumer, and a community wage for producing content; a user will become a content producer if the gain from doing so is seen as bigger than the cost. Bruns (2012) argues that the openness and simplicity of Twitter's platform has played an important role in its success thus far, but warns that a balance between the needs of platform providers, users and third-party developers is vital in retaining social media innovation and development. While Twitter remains a platform for sharing content and opinions, several important events over the past few years have shown its indisputable value as a communication channel during emergencies, perhaps especially so when mobile access to social media is the most effective communication alternative (Bruns 2012). Hughes and Palen (2009) argue in favor of using micro blogs as a public information channel used by authorities, for instance in emergency situations.

Naturally, an increase in commercial businesses actively pursuing Twitter as a means to reach consumers offers Twitter the chance to increase their advertising revenues, which in turn could be employed to enhance their product. However, Twitter's users have also contributed significantly to its development. Important features such as the ability to identify usernames using @, or identifying topics or keywords using hashtags (and later; cashtags) emanated from Twitter's user base (Bruns 2012; Madrigal 2013).

Although many news sites allow comments on their articles, there is demand by investment professionals for a common and independent place to discuss the news. "Traders and investors alike have come to view these platforms as trusted filters that help them make more informed decisions because they can discuss and interpret the news with their peers" (Zeledon 2009). Twitter is one of the platforms that can satisfy this demand.

The Predictability of Stock Markets

It is extremely desirable to be able to predict the stock market and a myriad of models have thus been developed for this purpose. Some of these models successfully predict returns for past data, but often fail in later attempts (Bodie, Kane, and Marcus 2011, 367). One possible explanation for this could be that the market adjusts for these new methods, so that they are no longer profitable. This discourages investors from sharing successful and potentially profitable models. Academics may be tempted to sell their work to the financial industry instead of publishing their work. For example, Prof. Johan Bollen, whose work we cite in our paper, teamed up with hedge fund manager Paul Hawtin to launch a hedge fund based on his algorithms (Kelly 2011).

Stock market analysis was one of the first applications of computers in economics. Maurice Kendall famously discovered in 1953 that "he could identify no predictable patterns in stock prices. Prices seemed to evolve randomly" (Bodie, Kane, and Marcus 2011, 343). Such findings eventually led to the development of the efficient market hypothesis which states that stocks already reflect all available information, making it impossible to predict their movement based on past data (Bodie, Kane, and Marcus 2011, 345). Others, such as Paul Tudor, a hedge fund manager and trader, believed that markets were showing repeated

patterns and received attention for his accurate prediction of the 1987 Black Monday stock market crash (Trejdify 2012).

While conventional financial theory usually assumes full rationality and efficiency, consensus in behavioral finance is that psychology and emotions are important factors in determining how investors behave. This may lead to deviations from market efficiency which is imperative for technical analysis and arbitrage strategies to work. Irrational investor behavior is seen as an opportunity by arbitrageurs. They take advantage of irrational behavior and make profits by taking opposite positions. Modern behavioral finance theory suggests that humans are not rational machines, rather emotional, rationally bounded, and subjective actors, who are influenced by things other than the cold facts (Bodie, Kane, and Marcus 2011, 356). Several studies find that psychology affects investor behavior. One example is a study that found "a significant market decline after soccer losses" (Edmans, Garcia, and Norli 2007). This shows that the mood of investors may influence the stock market.

There are currently two main types of algorithm trading; arbitrage robots seek to identify mispriced securities whereas high frequency algorithm trading is based on exploiting pricing errors and illiquidity in stocks. The latter makes profits by simultaneously offering to buy and sell stocks on both sides of the spread. Both are founded on market inefficiencies, and inefficient markets are therefore a prerequisite for them to be profitable. As the number of robots engaged in the market increases, it becomes harder for them to be profitable.

Big data analysis algorithms, on the other hand, exploit the inability of investors to consider all relevant information, and could be profitable even if the market is weak-form efficient. Unless insider information is leaked through social media, such analysis will not work if the market is semi-strong form efficient. Thus, our results will indicate a certain level of efficiency. Indeed, if we are able to predict the market, our study will add to the list of proof for market inefficiency.

Automated trading is not without problems. People have attempted and succeeded in tricking high frequency algorithm trading, including a famous case in Norway where two day-traders successfully profited from deceiving the robot Timber Hill.

The Supreme Court of Norway (2nd of May 2012) found them not guilty of market manipulation. This ruling affirmed that market participants are responsible for their own actions – fooling irrational robots should not be illegal. As long as there is money to gain, there will be incentives to trick big data algorithms as well. For example, this could be achieved by distributing false rumors.

Twitter's Influence on the Stock Markets

As mentioned before, investors are already using big data analysis to make investment decisions. However, this data has to come from somewhere. One of the pioneers in using Twitter for stock related chatter is StockTwits. It has been said that "StockTwits is the modern version of traders shouting in the pits" (Zeledon 2009). It is thus the traders themselves who are the data source in this case. The discourse about stocks can in itself be valuable information, as it amongst other things indicates investor interest. Indeed, a study of the noise level in trading pits found that it could actually be used to predict several aspects of the market, such as the volume of trades and volatility (Coval and Shumway 2001).

A recent trend we observe is that companies themselves are beginning to publish price-sensitive information such as earnings announcements on social media. Of course, such information is usually published through several channels simultaneously making it impossible to isolate the effect of one specific post. One prominent exception was when the CEO of Netflix posted a message on Facebook stating that they had passed 1 billion hours of viewings per month. The news first spread on social media before being picked up by mainstream news. This led the stock price to increase by 16% by the end of the next day (Scannell 2013). In such instances, those monitoring social media are clearly at an advantage. This message initiated an investigation by the Securities and Exchange Commission on whether the message breached disclosure regulations. One of the main points was that "the post was not accompanied by a press release, a post on Netflix's own web site or Facebook page, or a Form 8-K," meaning that this was new information spread to the market solely through an unorthodox source (SEC 2013a). On the 2nd of April 2013, the SEC determined not to pursue an enforcement action on the matter and approved social media such as Twitter and Facebook as valid communication channels for stock sensitive information as long as this is made clear in advance (SEC 2013b). As a consequence, we should expect more companies to embrace

Twitter as an official communications channel for stock sensitive information in the future.

One recent incident clearly demonstrates that Twitter not only influences individual companies, but also the market in general. On the 23rd of April 2013, hackers assumed control over Associates Press' Twitter account and posted the following message: "Breaking: Two Explosions in the White House and Barack Obama is injured" (Kisling, Lam, and Mehta 2013). Some of the response on Twitter to this tweet is displayed in Table 2. This triggered a 0.9 percent immediate decline in the S&P 500, as can be seen from the chart in Figure 2, wiping out about $136 billion in market value from the companies in the index. Even more extreme, the VIX (a volatility index, which correlates negatively with S&P 500 most of the time) surged more than nine percent in the two minutes after the tweet. The market recovered within three minutes as investors determined that the post was incorrect. Some traders said the dip might have been caused by algorithm trading robots tracking the news headlines, reacting contrary to humans, who would have most likely verified the information before trading on it (Kisling, Lam, and Mehta 2013). This incident will most certainly have consequences for algorithms employed in the market. They will be adjusted, so that such events are less likely to happen again. One way to do this is to require verification from a second original and trusted source before trading.

Table 2 - Some of the Tweets Following the 23rd of April AP Hack

Time	User	Tweet
13:09:02	DAK	wowmany machines $ES_F
13:09:02	Jason	wow $es_f whats going on?
13:09:02	Jordan	Woah $spy
13:09:23	Jeff C Brook	whats going on with the $NDX charade? #timestamp
13:09:38	Beautiful Kitty	What just nuked the market? $DIA
13:09:41	Michael J Zoitas	Damn skippy $es_f quick break down off the 1575 level
13:10:07	WiseRguy	$SPY $SPX what happened?
13:10:12	Large Void Bot	13:10 Drops: $XIV -5.3%, $SVXY -4.9%
13:10:17	Tim Trice	WOW! And that's what happens when you have a bull run on instability! $SPY $IWM
13:10:26	Lin/	yes,. $VXX to infinity and beyond
13:10:26	serge	chatter of explosions $SPY $TLT being heard
13:10:33	Jack Damn	What just happened? $SPY
13:10:42	W C Hsueh	Market suddenly drops .. Go figure $SPY $QQQ $DIA 1:00pm sell program?
13:10:55	Jon	WOW major intraday DUMP! $AAPL $SPY $IBB #captainobvious here to help!
13:10:57	WiseRguy	$SPX $SPY can't be cauz of this bs news "FAA says sequester-related furloughs Monday delayed 1,200 flights."
13:11:02	Alan Tu	What just happened to $USDJPY? Dropped 40 pips, I got out long an hour ago #forex #fxtalk
13:11:03	Berry Cobb	AP Hack and the Algos go wild! $ES_F etc....
13:11:11	DA_BARRON	JUST LOOKED LIKE SOMEONE FARTED IN A CROWDED ELEVATOR $SPX $COMP $INDU #NOSEDIVE
13:11:16	Leigh Drogen	Did a Twitter account hack just cause a big market dislocation, wow $SPY
13:11:30	Adam Tang	Right after I finished lunch too! $ES_F oh here we go the other way!
13:11:41	Day Traders Elite	What in the world just happened? Bad news? $SPY
13:11:53	dirty harry	AP was hacked. No one hurt.$SDS $SPY $VXX via BUZZFEED

Figure 2 - Drop in the S&P 500 due to a Single Tweet

Source: Yahoo! Finance

Another recent example of the influence tweets can have on stocks is when well-known investor Carl Icahn tweeted that he had a large share of Apple stock, and that he had been in talks with the CEO of Apple, Tim Cook. The stock surged almost USD 5 in just 4 minutes after the tweet, implying a gain of over 4 billion USD in market cap for Apple due to a single tweet.

Figure 3 - Reaction to Tweet from Carl Icahn

Source: Yahoo! Finance

All stocks are not equally appropriate for Twitter analysis. While unknown companies rarely are discussed on Twitter, the opposite problem arises with large companies where wide discussion and rigorous analyst coverage seems to make them too efficiently priced for such analysis to be effective. Baker and Wurgler (2006) argue that small stocks, which are difficult to arbitrage, are more likely to

be affected by emotion rather than highly liquid stocks. They hypothesize and find that "investor sentiment has larger effects on securities whose valuations are highly subjective and difficult to arbitrage". Small stocks are also more likely mispriced because they are often ignored by analysts: "Small stocks that receive relatively little coverage by Wall Street analysts may be less efficiently priced than large ones" (Bodie, Kane, and Marcus 2011, 346). This implies that studies on the stock level could benefit by including at least two different groups of stocks (large and small) and compare the differences. However, when examining our data set, we realized that such analysis would require a much longer time period due to the limited number of daily tweets mentioning individual companies. Therefore we only consider aggregate sentiment measures and stock index returns in this thesis.

One of the primary factors enabling big data analysis is the rapid development of information technology. In particular, the increase in computational power has been exponential within the last decades with capacity remarkably closely following Moore's law, doubling every 18 months (Kanellos 2003). If this development continues, computers will eventually outperform humans in big data analysis as they have done in other areas where humans traditionally have had the edge, like checkers, chess, Jeopardy! and Scrabble. For example, the first time a computer beat a top human being in chess was February 10, 1996 (IBM 2012). The last time a top human beat a top computer in chess was on November 21, 2005, and this will probably never happen again. Indeed, even modern mobile phones are now able to reach grandmaster level in chess (Ramos and Islam 2012).

Sentiment Analysis

There are two main approaches of conducting sentiment analysis, depending on the level of supervision researchers choose to adopt (Ghiassi, Skinner, and Zimbra 2013). Researchers can either perform the analysis using unsupervised or supervised analysis. In unsupervised analysis the text material's sentiment is determined using statistical techniques, algorithms or lexicons containing positive and negative terms (Redmore 2012). Several commercial tools may be applied to perform unsupervised analysis to classify sentiment, or researchers may create such a tool themselves using Excel, for example. Alternatively, researchers can choose to supervise their analysis using machine learning algorithms. Using the

supervised approach is more laborious, but has the potential to yield better sentiment classification accuracy (Pang, Lee, and Vaithyanathan 2002; Sebastiani 2002). There are several steps in preparing and developing machine learning algorithms in order to perform supervised analysis. First, researchers have to collect a corpus of text data (tweets, in our case) and process the data to prepare it for analysis using for instance natural language processing (NLP) techniques. Different methods can then be employed to classify data; using Parts-of-Speech, N-grams and machine learning classifiers such as Naïve Bayes, Support Vector Machines (SVMs) and Maximum Entropy, among others (Redmore 2012; Go, Bhayani, and Huang 2009), either separately or combined. After developing the classifiers, researchers need to train them in order to evaluate and improve their effectiveness (Sebastiani 2002). This is normally achieved by manually classifying the sentiment of a subset of the obtained data set – obviously a laborious and time-consuming task, followed by training the classifiers on the subset, and finally analyzing data using the training classifiers.

In determining the sentiment, the factors that induce people to use positive and negative wording have to be considered. One study found that the specific words people use in tweets are not only related to their opinion of whether to buy or sell a certain stock, but also dependent on the general mood: "people start using more emotional words such as hope, fear and worry in times of economic uncertainty, independent of whether they have a positive or negative context" (Zhang, Fuehres, and Gloor 2011). Thus, volatile periods could be predicted by measuring the amount of emotional words. However, performing such a study would require data including at least one crisis to perform a cross section study, data we do not have.

The Forecasting Power of Twitter

Our focus is on finding out whether Twitter can be used to predict stock prices. There are a few studies on the predicting power of Twitter. Most of them, however, are not finance related. Nevertheless, we include some of them due to the rarity of such studies.

A paper by Asur and Huberman (2010) demonstrates how Twitter data can be used to forecast box-office revenues for movies. They found that "a simple model

built from the rate at which tweets are created about particular topics can outperform market-based predictors". Another study found that "the volume of blog posts about an album is positively correlated with future sales" (Dhar and Chang 2007). It is unclear to what degree social media is serving as a proxy for existing market interest or actually in itself leads to increased publicity and sales. The main point, however, is that both studies are successfully able to use social media to forecast sales.

One highly relevant study investigates "whether measurements of collective mood states derived from large-scale Twitter feeds are correlated to the value of the Dow Jones Industrial Average (DJIA) over time" (Bollen, Mao, and Zeng 2011). They used the popular mood-tracking tool OpinionFinder as well as their own tool, GPOMS, and found predicting power for some of the public mood dimensions. OpinionFinder did not prove to be particularly effective in predicting the DJIA, but the GPOMS dimension "calmness" was a good predictor. Contrary to most other similar studies, this research has actually been applied in a real world hedge fund. As mentioned above, Paul Hawtin of Derwent Capital Markets collaborated with Johan Bollen to launch a fund which would make daily investment decisions based on the output from the model. However, the fund was shut down after just one month, supposedly to develop an online trading platform. Reportedly, the fund did actually work and returned 1.86 percent ahead of the market and average hedge fund (Bloomberg 2013a).

A study by Sprenger and Welpe (2010) is also closely related to ours. They found "the sentiment (i.e., bullishness) of tweets to be associated with abnormal stock returns and message volume to predict next-day trading volume". The study focuses on stock specific tweets to predict aspects of four major US market indices. They utilize the concept of cashtags, created by putting the dollar sign before the ticker (e.g. $AAPL), as search strings to filter for stock relevant tweets. Cashtags were initially created by stocktwits.com, but were later officially adopted as standard modifiers by Twitter (Scannell 2013). Cashtags have greatly increased in popularity since then, and we therefore expect to obtain a much larger daily volume of tweets than Sprenger and Welpe, also likely representing a greater diversity of market participants. Sprenger and Welpe's study has some interesting and motivating results. They found that tweet volume predicted trading volume

and tweet bullishness predicted abnormal returns. We will adopt several aspects of their methodology for our thesis. We adopt the use of cashtags to identify stock related tweets. However, instead of only including the stocks included in the S&P 100, we include all stocks, giving us a much broader sample to determine overall investor sentiment.

In her BI Norwegian Business School master thesis, Jubbega (2011) found that brand sentiment tweets had an effect on the stock price for 5 of 10 companies. She found that investor reactions grow over time, peaking after 2 to 4 days, then decline 1 to 6 days after the peak. Our study differs in that it takes a finance rather than marketing perspective. Also, instead of using the mentioning of brands, we use mentioning of stock tickers. This provides us with a data set of tweets specifically concerning the stock, instead of general discourse about the company or its brands.

Summary of Literature Review

Several previous studies have shown that Twitter data can be used as a leading indicator in a wide range of settings, including the stock markets. The extent of its power, however, is still unclear and requires further research. And although there is some evidence that such analysis is used in actual trading, the success of such trading has so far not been well covered in academic research. Very little research has been conducted examining the stock specific discourse on Twitter.

Hypotheses

Although causal relationships between tweets and stock returns have been indicated in previous research, most studies note that their models have limitations and should be retested in further research. Also, few studies show any diagnostic tests of their models.

We want to investigate whether stock specific tweets can be used to predict stock returns. Some of the earlier studies have used the entire firehose from Twitter. If we can achieve the same or even better results with a more relevant subset of tweets, it will dramatically reduce the cost and need of computational power.

Our data set is unique. Although Sprenger and Welpe (2010) also used cashtags, it was done at a point in time when they were not yet widely used. Today, cashtags have been adopted by a wide range of participants in the financial industry. While Sprenger and Welpe collected less than a quarter million tweets over a six-month period (although not directly comparable since they limited themselves to S&P 100 stocks) we collected over two and a half million tweets in four months.

Previous studies have shown that both positive and negative sentiment can predict stock returns. Emotions in general have also been found to predict stock returns. Nevertheless, our null hypothesis is that the efficient market hypothesis is true, implying that stock prices cannot be predicted based on past information.

We thus present the following hypotheses:

$H1_0$:   Positive sentiment does not predict stock index returns

$H1_A$:   Positive sentiment predicts stock index returns

$H2_0$:   Negative sentiment does not predict stock index returns

$H2_A$:   Negative sentiment predicts stock index returns

$H3_0$:   Emotional tweets do not predict stock index returns

$H3_A$:   Emotional tweets predict stock index returns

Data

We have chosen to focus exclusively on the US market mainly because it is currently Twitter's largest market by far. In fact, over 30% of tweets are originating from within the US (Wrenn 2012). Particularly, discourse specific to stocks seems to be mostly a US phenomenon for now. Our initial investigations revealed that about 84% of all tweets containing a stock ticker were written in English. Most of the noise also seemed to come from non-English languages. We therefore excluded any non-English tweets from our sample.

Due to the limited amount of tweets for individual stocks, except a few highly discussed stocks (such as Apple, Citibank, and Microsoft), we only consider the overall market. Analyzing the effects on individual stocks is possible, but would require a much longer time period.

Data Collection

Because Twitter no longer provide free access to the firehose of all tweets, we have to rely on a third party commercial service to obtain our tweet data. We have chosen to use DataSift which allows us to filter and download relevant tweets from the entire Twitter firehose. Following Sprenger and Welpe (2010), we use cashtags to identify stock specific tweets in our initial filter. Using only tweets that contain cashtags will not give us all tweets relevant to a given company, but we believe this is the best approach to obtain a useable data set for our study. In addition we filter for tweets written in the English language. This limits our data set and reduces noise.

Although almost every tweet containing $ followed by a letter is stock related, there are some exceptions, such as replacing S with $ (e.g. ca$h). We assume this is random noise which should not affect our results. Searching specifically for the ticker of each company would have given us a cleaner data set, but would also have been much more costly. Therefore, we chose to search for stock related tweets using the dollar sign followed by each letter of the English alphabet. An excerpt from our search syntax can be seen in Figure 4.

Figure 4 - Search Syntax

```
language.tag in "en"
AND
(twitter.text substr " $A"
OR
twitter.text substr " $B"
OR
twitter.text substr " $C"
OR … )
```

By collecting only tweets mentioning stock tickers, we have an effective way of limiting our data set to stock related tweets, excluding much of the noise and ambiguous meanings generally found on Twitter. For example, we would not have been able to review companies such as Apple if we searched by company names as that would include a lot of fruit related tweets hardly relevant for our thesis. Also, to analyze all tweets to obtain overall sentiment has become unrealistic for most researchers due to the extreme volumes involved. While the total daily volume is over 400,000,000 tweets, only about 10,000 - 30,000 of them include a stock ticker, as can be seen from our data.

Some companies have started using cashtags in tweets relevant to their stock price. Table 3 shows an example of a company using Twitter to announce their earnings release. Notice the cashtag at the end. Also, the tweet is neutral and thus has a sentiment score of 0.

Table 3 - Earnings Tweet Example

Time and Date	Username	Followers	Sentiment
23.04.13 07:30 EST	Lockheed Martin	42000	0
Just released: First quarter 2013 results: http://t.co/9fRL3s6BFO $LMT			

Properties of the Data

Most stock related talk on Twitter happens, as expected, during US trading hours. As can be seen from the histogram in Figure 5, the highest tweet frequency occurs at the opening of the exchanges. There is also a spike at closing leading to a U-shape during trading hours. This reflects previous research on trader discourse activity. Coval and Shumway (2001) recorded the ambient noise levels in physical trader pits and found that the sound level was highest on opening and closing of the exchange. We observe, however, that the average end of day spike happens after the exchange has closed at 4 p.m. We believe this is because the histogram is

based partly on data from the earnings season and many companies report their earnings at 4 p.m. The release of earnings naturally generates discussion on Twitter. Other explanations can be after-hours trading or discussion of trades already done.

Figure 5 - Aggregate Histogram of Daily Tweet Volume

This histogram shows aggregate daily tweet volume in 10-minute intervals for the entire data set. The darker area indicates the opening hours of the exchanges.
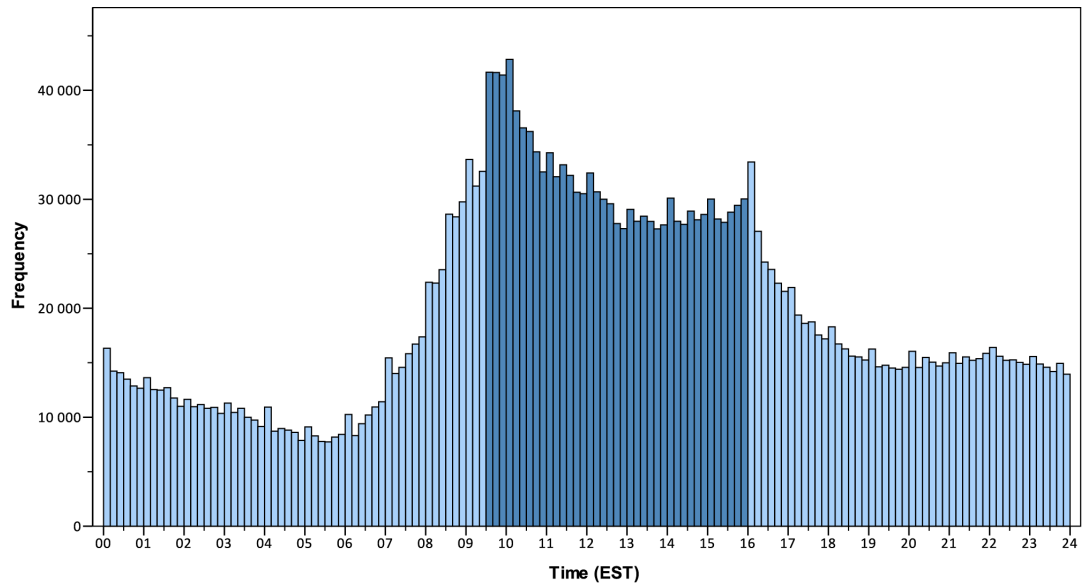
From a day-to-day perspective, we see some variation in activity, but the pattern is very similar. In Figure 6, we see the activity during a typical week. We observe that activity is highest during the operating hours of the exchanges and generally low on weekends.

Figure 6 - Weekly Histogram of Tweet Volume
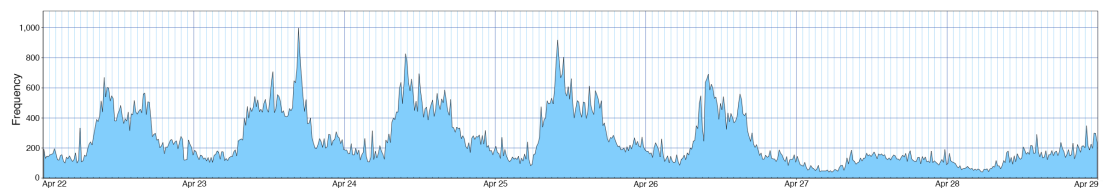
In Figure 7 below, we have summed the number of tweets from 4 p.m. a given day until 4 p.m. the next day. We see, as expected, that stock related message volume is much lower on the weekends than on trading days.

Figure 7 - Tweet Volume per Day from Week 17 to Week 32

The diagram shows the stacked distribution of tweets for each weekday from 4 p.m. the preceding day until 4 p.m. the indicated day. Note that this diagram does not include the first three days, and the last five days of our data set, because we are comparing whole weeks only (Monday-Sunday).
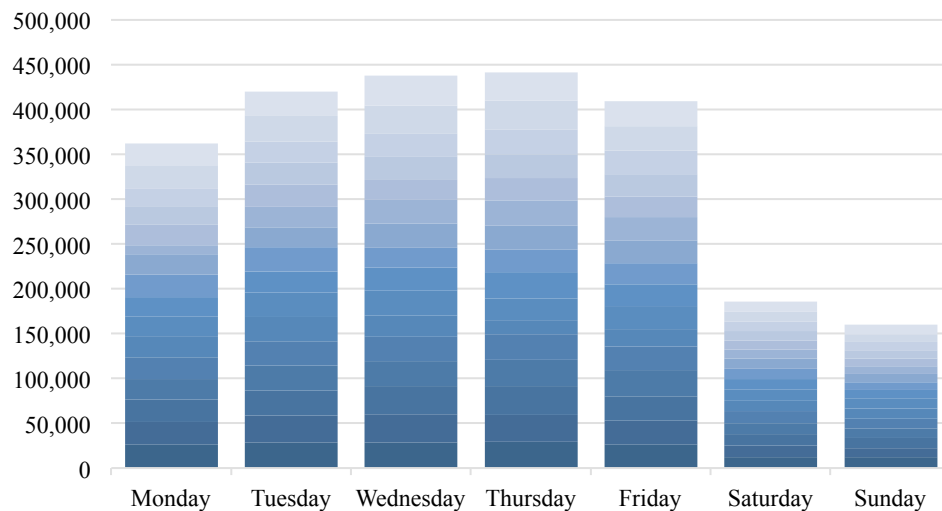
Sentiment Measure

A common approach to testing the predictable power of Twitter is tracking the sentiment of tweets. The sentiment of tweets can be viewed as a proxy for the general mood in the market, which as shown in the literature review can affect stock prices.

We will not attempt to create a sentiment measure superior to the professional solutions available, instead we have chosen to rely on a solution provided by DataSift. They use Lexalytics' Salience Engine to calculate the sentiment measure (DataSift 2013; Stenson 2012). The Salience Engine uses natural language processing and supervised machine learning techniques to return a fine-grained sentiment scale scoring each tweet from -100 to 100 in sentiment, where -100 is most negative, 100 is most positive, and 0 indicates no sentiment measured. While not always adapted, the ability to determine sentiment degree could be advantageous for some studies. However, such a fine granularity might lead to discrepancies when humans are validating the computer-generated score since human perception of sentiment is subjective. (Ghiassi, Skinner, and Zimbra 2013).

As we do not know the magnitude of DataSift's sentiment measure, and therefore cannot quantify the level of negativity or positivity, we generalize the tweet

sentiment score as positive, negative or neutral. Also, we aggregate the measure to the daily level to match our stock return data.

Previous studies have shown, however, that binary sentiment measures have inferior predictable powers on the market (Bollen, Mao, and Zeng 2011). The human mood is much too complex for a simple separation into positive, negative and neutral mood states, which many sentiment algorithms are based on. Future studies should therefore strive to use more advanced sentiment classification logarithms. Bollen, Mao and Zeng found the strongest support for calmness, while Zhang, Fuehres, and Gloor (2011) found strong support when measuring hope, worry, fear, anxiousness, and negativity in general. Interestingly, both positive and negative words were found to be negatively correlated with market returns for all four exchanges examined yet positively correlated with the VIX. This indicates that emotional outburst in general can be used to predict poor stock performance. We test separately for this by creating an additional variable by adding up both positive and negative tweets divided by the total number of tweets for each day.

Consistent with findings in previous research, our data set contains more positive than negative tweets. In our sample, 21% of tweets are identified as positive while 12% are identified as negative. The remaining 67% of tweets are classified as neutral. We separate positive and negative sentiment as we expect the effects to be asymmetric. One reason for this is that people are conditioned through advertising to not trust positive information. Companies are very good at highlighting the positive sides of their businesses while hiding the negative sides. As a substantial amount of tweets have a commercial purpose, we expected to find more positive than negative tweets. Negative signals may however carry more weight because of their higher inherent credibility.

Sample Size

Following Sprenger and Welpe (2010), we set the daily cutoff to be concurrent with markets closing at 4 p.m. EST since all tweets after this point can only affect the next trading day. We achieve this by adding 4 hours to the GMT time stamp, so midnight is defined as 4 p.m. EST. Thus each day is defined as the time between each closing of the stock markets in the US.

We have collected data continuously since the 18th of April at 4 p.m. GMT. Since we set the daily cutoff at 4 p.m. EST, our first data entry is at 4 p.m. EST on the 18th of April. This data belongs to the 19th of April trading day, so that is the first day in our data set. We exported our data set on the 16th of August, and thus have 120 days of data.

Over the four-month period, we collected a total of 2,842,248 tweets. Of these, 2,599,277 (91.45%) were assigned a numerical value for sentiment. We only include tweets with a sentiment value in our data set. To deal with the missing stock prices on weekends, we assign all tweets after 4 p.m. on Fridays to the measure for Monday. In this period there are two trading holidays, on the 27th of May and the 4th of July. Tweets sent during one of these days are also assigned to the next trading day. This gives us 84 trading days. We obtain the return data of the S&P 500 for these trading days from Yahoo! Finance.

Variables

Our variables are described in Table 4 below. We calculate returns (RET) from the daily S&P 500 closing values obtained from Yahoo! Finance. The measure for positive sentiment (POS) is created by summing up the number of tweets containing positive sentiment on a given day and dividing this number by the total number of tweets that day. The variable thus represents the percentage of tweets with a positive sentiment on a given day. The negative sentiment variable (NEG) is generated in the same way. We create the emotionality (EMO) variable by adding both the positive and negative tweets and divide by the total number of tweets. H is a dummy variable indicating days after non-trading days such as holidays and weekends.

Table 4 - Variable Descriptions

RET	Stock index returns of the S&P 500
POS	Percentage of positive tweets
NEG	Percentage of negative tweets
EMO	Percentage of tweets which are either positive or negative
H	The day after holidays (dummy variable)

The reason we do not use the number of positive or negative tweets as variables is because the total tweet volume varies by day. More importantly, since we

aggregate all tweets over the weekend to the Monday measure, we have substantially more tweets for Mondays compared to other days in the week. By taking the relative number of positive or negative tweets compared to the daily total, we eliminate most of this problem. However, we still have reason to believe days after holidays are significantly different to other days. Our investigation indicates less emotionality on weekends and holidays, causing Mondays to have a lower percentage of emotional tweets, as can be seen in Figure 8. This could be because a part of the emotions in stock related tweets originates from changes in stock prices. Thus we feel it is right to add a dummy variable to account for this difference.

Figure 8 - Percentage of Emotional Tweets per Day
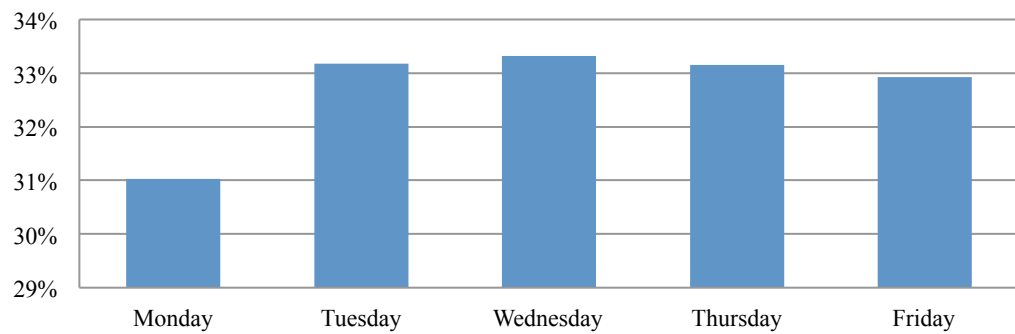
Descriptive Statistics and Correlations

Table 5 below shows the descriptive statistics of our variables. We see that RET has a mean close to zero, as expected, while the others have a positive mean. The mean of EMO is by definition the sum of POS and NEG.

Table 5 - Descriptive Statistics

	RET	POS	NEG	EMO	H
Mean	0.0009	0.2058	0.1217	0.3276	0.2143
Median	0.0016	0.2041	0.1226	0.3285	0
Maximum	0.0148	0.2404	0.1369	0.3625	1
Minimum	-0.0250	0.1746	0.1020	0.2879	0
Std. Dev.	0.0073	0.0124	0.0084	0.0167	0.4128
Skewness	-0.6801	0.4675	-0.3744	-0.0559	1.3926
Kurtosis	3.8245	3.3666	2.5100	2.4528	2.9394
Jarque-Bera	8.8553	3.5300	2.8032	1.0919	27.164
Probability	0.0119	0.1712	0.2462	0.5793	0.0000
Sum	0.0737	17.290	10.224	27.515	18
Sum Sq. Dev.	0.0045	0.0128	0.0059	0.0233	14.143
Observations	84	84	84	84	84

We observe that RET failed the Jarque-Bera normality test. As can be seen from the histogram in Figure 9, this is likely caused by an outlier. We will consider removing the outlier if non-normality becomes a major problem in our models.

Figure 9 - Histogram of RET
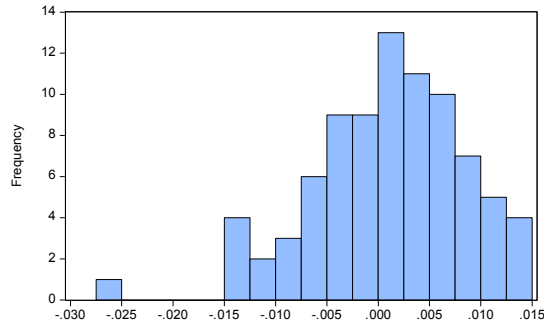
Correlations among our variables are displayed in Table 6 below. Interestingly, there seems to be an increase in the share of tweets assigned sentiment during the period, as POS, NEG and EMO are all positively correlated with the date. This may indicate that tweets have become more emotional over time or that the sentiment measure has changed. This could potentially create a problem with heteroskedasticity and non-stationarity.

Of course, positivity (POS) and negativity (NEG) is strongly positively correlated with emotionality (EMO). Returns (RET) are also not surprisingly negatively correlated with negative tweets. However, RET seems to be uncorrelated with POS. Because EMO is strongly correlated with POS and NEG, we should not include all of them in one model, as this would likely give us a problem with multicollinearity. Therefore we create one model with POS and NEG as the explanatory variables and a second with EMO as the explanatory variable.

Table 6 - Correlation Matrix

	DATE	RET	POS	NEG	EMO	H
DATE	1	-0.16	0.35	0.44	0.48	0.00
RET	-0.16	1	-0.02	-0.26	-0.15	0.09
POS	0.35	-0.02	1	0.26	0.87	-0.46
NEG	0.44	-0.26	0.26	1	0.70	-0.37
EMO	0.48	-0.15	0.87	0.70	1	-0.53
H	0.00	0.09	-0.46	-0.37	-0.53	1

Methodology

In this section we will outline the statistical tests and methods we will use to test our hypotheses.

We have chosen to employ the Vector Autoregressive Regression (VAR) approach for several reasons: it is appropriate for time series; allows us to have several endogenous variables and lets us investigate feedback effects. A VAR allows the value of a variable to depend on both its own lags and lags of the other variables in the system. In addition, a VAR also allows for dummy variables as exogenous variables to account for structural changes in our data.

We believe a simultaneous equations model is necessary since tweets are as likely to be affected by stock returns as stock returns are to be affected by tweets. This suggests that all our variables except the dummy variable should be treated as endogenous. Investor sentiment and stock returns ought to be simultaneously related since happy investors may buy more stocks, and rising stock prices tend to improve the mood of investors.

Model 1 - Positive and Negative Tweets

Below is a mathematical representation of our first VAR model in matrix form:

$$\begin{bmatrix} POS_t \\ NEG_t \\ RET_t \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \sum_{i=1}^{k} \begin{bmatrix} \beta_{11}^i & \beta_{12}^i & \beta_{13}^i \\ \beta_{21}^i & \beta_{22}^i & \beta_{23}^i \\ \beta_{31}^i & \beta_{32}^i & \beta_{33}^i \end{bmatrix} \begin{bmatrix} POS_{t-i} \\ NEG_{t-i} \\ RET_{t-i} \end{bmatrix} + \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{bmatrix} H_t + \begin{bmatrix} u_{POS_t} \\ u_{NEG_t} \\ u_{RET_t} \end{bmatrix} \quad (1)$$

Where *t* is time, the α's are the intercepts, *k* is the number of lags, *i* is the lag number and *u* are white noise disturbance terms which are assumed to be uncorrelated with each other and have zero mean. The variables are described in the previous section.

Model 2 - Emotional Tweets

In the same fashion, below is a representation of our second VAR model, investigating the relationship between returns and emotionality in general:

$$\begin{bmatrix} EMO_t \\ RET_t \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} + \Sigma_{i=1}^{k} \begin{bmatrix} \beta_{11}^i & \beta_{12}^i \\ \beta_{21}^i & \beta_{22}^i \end{bmatrix} \begin{bmatrix} EMO_{t-i} \\ RET_{t-i} \end{bmatrix} + \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} H_t + \begin{bmatrix} u_{EMO\,t} \\ u_{RET\,t} \end{bmatrix} \quad (2)$$

Diagnostic Tests

All variables in a VAR have to be stationary for the estimations to be valid. Intuitively, we see no reason our variables should not be stationary. We measure relative positivity and negativity which should not be expected to follow any specific trend. Also, even though stock prices are usually non-stationary, returns should be random and stationary. Nevertheless, we test all our variables for stationarity using the Augmented Dickey-Fuller (ADF)-test.

We will test for autocorrelation in the residuals by using the Portmanteau test of autocorrelation. Because stock return volatility tends to vary over time, we expect to encounter some degree of heteroskedasticity. For example, our data includes both the April-May and July-August earnings seasons, as well as the summer holidays. We use White's test on the residuals to check whether our variables are heteroskedastic. Finally, we will test whether the residuals are normally distributed using the Jarque-Bera test.

Tests of the Hypotheses

We will test our hypotheses using Granger causality tests, impulse response functions (IRF), and variance decompositions. These methods are explained when we apply them on our models in the next section. Throughout, we will use a significance level of 5%.

Results

Here we present the results of our methodological tests, and review the consequences these results have for the validity of our hypotheses.

Stationarity Tests

Inspecting the plots of our variables over time, we see no sign of any trends. We use the ADF test for unit root to formally check whether our variables are stationary. We use the Schwarz Information Criterion to determine the appropriate lag length. RET has a mean close to zero, so we do not include an intercept for that variable. The sentiment variables have a non-zero mean, and we thus include an intercept for them. As can be concluded from the results summarized in Table 7 below, non-stationarity is rejected at the 5% level for all variables. We therefore assume that we do not have a problem with non-stationarity. Since our variables are stationary, there is no need to run cointegration tests.

Table 7 - Augmented Dickey-Fuller Unit Root Test Results

Variable	p-value	Constant	Lag length[1]
RET	0.00	No	0
POS	0.00	Yes	0
NEG	0.02	Yes	1
EMO	0.00	Yes	0

[1] Determined by the Schwarz Information Criterion

Number of Lags

We need to determine the optimum number of lags for our model. In EViews, we can simultaneously test for several of the most popular measures. In the test results shown in Table 8, we see that for model 1, the Likelihood Ratio (LR) test, the final prediction error (FPE) and the Akaike information criterion (AIC) indicate that two lags are optimal whereas the Schwarz information criterion (SC) and Hannah-Quinn (HQ) indicate one lag. For model 2, LR, FPE and AIC indicate 4 lags whereas SC and HQ indicate one lag. Based on these results, we choose two lags for model 1. We first estimated model 2 with 4 lags, but since it seemed to have serious problems with autocorrelation we chose to use one lag instead.

Table 8 - Optimum Lag Length

Model	LR	FPE	AIC	SC	HQ
1	2	2	2	1	1
2	4	4	4	1	1

We may now rewrite our regression equations with a specific number of lags.

Model 1:

$$POS_t = \alpha_1 + \beta_{11}^1 POS_{t-1} + \beta_{11}^2 POS_{t-2} + \beta_{12}^1 NEG_{t-1} + \beta_{12}^2 NEG_{t-2} + \beta_{13}^1 RET_{t-1} + \beta_{13}^2 RET_{t-2} + \gamma_1 H_t + \varepsilon_{POS_t} \quad (3)$$

$$NEG_t = \alpha_2 + \beta_{21}^1 POS_{t-1} + \beta_{21}^2 POS_{t-2} + \beta_{22}^1 NEG_{t-1} + \beta_{22}^2 NEG_{t-2} + \beta_{23}^1 RET_{t-1} + \beta_{23}^2 RET_{t-2} + \gamma_2 H_t + \varepsilon_{POS_t} \quad (4)$$

$$RET_t = \alpha_3 + \beta_{31}^1 POS_{t-1} + \beta_{31}^2 POS_{t-2} + \beta_{32}^1 NEG_{t-1} + \beta_{32}^2 NEG_{t-2} + \beta_{33}^1 RET_{t-1} + \beta_{33}^2 RET_{t-2} + \gamma_3 H_t + \varepsilon_{POS_t} \quad (5)$$

Model 2:

$$EMO_t = \alpha_1 + \beta_{11}^1 EMO_{t-1} + \beta_{12}^1 RET_{t-1} + \gamma_1 H_t + \varepsilon_{POS_t} \quad (6)$$

$$RET_t = \alpha_2 + \beta_{21}^1 EMO_{t-1} + \beta_{22}^1 RET_{t-1} + \gamma_2 H_t + \varepsilon_{POS_t} \quad (7)$$

AR Roots

Since our variables are stationary, we expect the VARs to be stationary as well. To confirm this we can inspect graphs of the AR roots. The VAR is stationary (stable) if all inverse roots of the AR polynomial have modulus less than one and lie inside the unit circle. This is necessary for the IRFs to be valid. The AR roots graphs for both models are shown in Figure 10 below, and we see that all roots are well within the circles, confirming that the VARs are stable.

Figure 10 - AR Roots Graphs for Model 1 (left) and Model 2 (right)
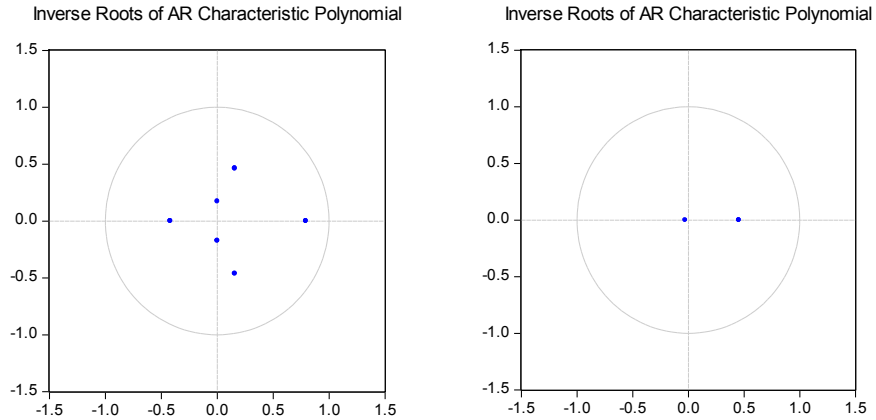
Diagnostic Tests on Residuals

The results of the diagnostic tests on our VAR model residuals are summarized in Table 9 below:

Table 9 - Results of Diagnostic Tests on Residuals

Model	Autocorrelation	Heteroskedasticity[1]	Normality[2]
(1)	No	0.288	0.003[*]
(2)	No	0.481	0.074

[*] null hypothesis rejected at the 5% level
[1] p-value of the joint White's test
[2] p-value of the joint Jarque-Bera test

In model 1, none of the orders of the Portmanteau autocorrelation test are significant, so we cannot reject the null hypothesis that there is no autocorrelation. Although model 2 has a significant autocorrelation at the 5% level at order 5, we conclude that autocorrelation is not an issue in this model either since all the other orders are insignificant.

According to White's general test of heteroskedasticity without cross-terms, we cannot reject the null hypothesis that the variance of the disturbance term is constant for both models. Thus we assume that we do not have a problem with heteroskedasticity.

The Jarque-Bera normality (Cholesky of covariance) test suggests that we must reject the null hypothesis that the residuals are normally distributed for model 1. However, inspection of the residuals in a Q-Q plot (Figure 11) reveals no serious deviation from normality. Using dummy variables to remove outliers might

improve normality, but has the downside of potentially being perceived as data mining since removing outliers will always reduce standard errors and increase R^2 (Brooks 2008, 166). Also, failure to meet the normality condition should not be of high concern for sufficiently large sample sizes (Brooks 2008, 164). With this in mind we have decided against taking action to obtain a higher degree of normality.

Figure 11 - Q-Q Plot of Residuals in Model 1
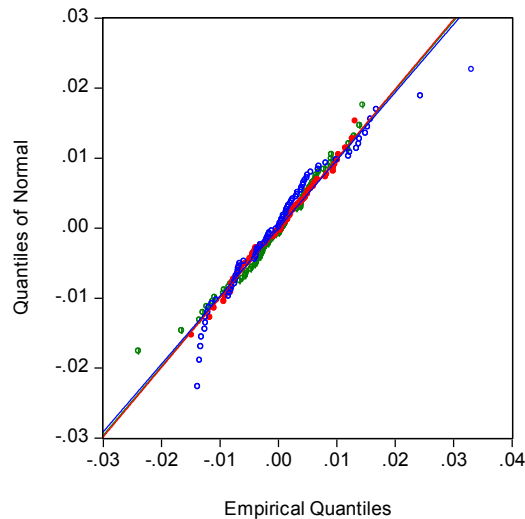
The models pass all diagnostic tests except normality for model 1. Nevertheless, as normality is not crucial, we continue to estimate the VAR coefficients, run causality tests, and estimate the IRF and variance decompositions for both models.

VAR Regression Output

From the VAR regression output in Table 10 we see that neither POS nor NEG has any significant effect on RET. For equation (5), R^2 is 9.4%, while adjusted R^2 is less than 1%, indicating a poor fit.

Table 10 - VAR Regression Output for Model 1

The table reports the VAR regression estimates for our first model, with t-stats in parenthesis.

	POS_t	NEG_t	RET_t
POS_{t-1}	0.324[*]	-0.084	-0.113
	(3.18)	(-1.23)	(-1.43)
POS_{t-2}	-0.080	0.109	0.024
	(-0.76)	(1.53)	(0.29)
NEG_{t-1}	0.203	0.391[*]	-0.176
	(1.29)	(3.67)	(-1.44)
NEG_{t-2}	0.297[*]	0.262[*]	0.001
	(2.06)	(2.69)	(0.01)
RET_{t-1}	0.166	-0.004	-0.027
	(1.06)	(-0.04)	(-0.22)
RET_{t-2}	0.314[*]	-0.093	-0.094
	(2.03)	(-0.89)	(-0.78)
C_t	0.098[*]	0.040[*]	0.040
	(3.75)	(2.25)	(1.98)
H_t	-0.013[*]	-0.009[*]	0.002
	(-4.61)	(-4.78)	(1.06)
R^2	0.422	0.478	0.094
Adj R^2	0.367	0.428	0.008
ser	0.009	0.006	0.007

[*] Significance at the 5% level

From the VAR regression output in Table 11 we see that EMO lagged one period has a significant negative effect on RET. EMO lagged one period has also a significant positive effect on itself. The R^2 is low for equation (7) as well.

Table 11 - VAR Regression Output for Model 2

t-stat in parenthesis

	EMO_t	RET_t
EMO_{t-1}	0.446[*]	-0.122[*]
	(5.63)	(-2.57)
RET_{t-1}	-0.029	-0.021
	(-0.16)	(-0.20)
C_t	0.187[*]	0.040[*]
	(7.18)	(2.59)
H_t	-0.022[*]	0.002
	(7.04)	(1.02)
R^2	0.502	0.086
Adj. R^2	0.483	0.052
ser	0.012	0.007

[*] Significance at the 5% level

Causality Test

The Granger causality test seeks to answer whether changes in one variable cause changes in another. However, since Granger causality is really only a correlation between the current value of one variable and the past values of itself and others, we cannot strictly say that one causes the other. What we can say, however, is that one variable tends to lead another. Table 12 shows the results of the Granger causality tests for both models. We observe a significant lagged relationship from NEG to POS, meaning that NEG Granger causes POS. Using emotional tweets as the independent variable, we see that EMO Granger causes RET.

Table 12 - Granger Causality Test Results

This table summarizes the Granger causality tests for both models. The test checks whether including the lags of another variable in addition to the lags of the dependent variable improves the forecast. The test simultaneously tests for all lags. The Chi-square values are reported, and p-values are shown in parenthesis.

Dependent	RET			POS		NEG		EMO
Independent	POS	NEG	EMO	NEG	RET	POS	RET	RET
	2.240 (0.326)	2.374 (0.305)	6.618* (0.010)	9.038* (0.011)	4.926 (0.085)	2.654 (0.265)	0.790 (0.674)	0.026 (0.872)

* Significance at the 5% level

It is important to note that "the lack of a Granger-causal relationship from one group of variables to the remaining variables cannot necessarily be interpreted as lack of a cause and effect relationship" (Lütkepohl 2006, 48). Thus, the IRFs can show relationships at individual lags, even though the combination of lags does not indicate any relationships.

Causality Diagrams

To summarize, the diagrams in Figure 12 and Figure 13 below illustrates the effects revealed in the Granger causality tests. The arrows indicate significant Granger causal relationships.
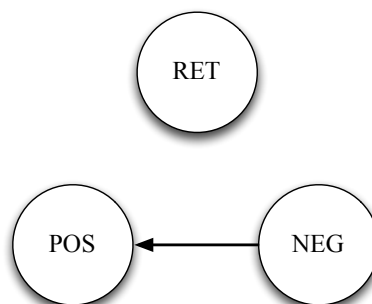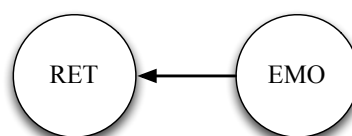
Figure 12 - Causality Diagram Model 1

Figure 13 - Causality Diagram Model 2

Impulse Response

One major issue with the Granger causality test is that it does not indicate any directionality of the effect (i.e. we cannot infer from the test results whether the effect EMO has on RET is negative or positive). The results also say nothing about how much time it takes for the effect of a shock in one variable to materialize in another. For this we use impulse response function (IRF) graphs. IRF graphs show the effect that an innovation (i.e. exogenous shock) in one variable has on some or all of the other variables over time. Analysis of the IRFs may be used to identify causal relationships: "if there is a reaction of one variable to an impulse in another variable we may call the latter causal for the former" (Lütkepohl 2006, 51). If the system is stable (stationary), the effects of the shocks should not be persistent (i.e. they should gradually fade away). We apply a shock of one standard deviation (because the variables have different scales) immediate change to each variable, and observe the response in the other variables over time. We only show the first seven lags in our tables because little interesting happens beyond that.

In Table 13 below we have summarized the results from the impulse response tests for all variables. The response of the variables to a shock in their own variable is of course very strong, and is not reported here.

Table 13 - Summary of Impulse Responses

This table summarizes the response from each variable to a one standard deviation shock in another. P-values are shown in parenthesis.

Dependent		RET			POS		NEG		EMO
Independent		POS	NEG	EMO	NEG	RET	POS	RET	RET
Lag	1	0.0018*	-0.0014	0.0005	0.0000	0.0000	-0.0005	0.0000	0.0000
		(0.025)	(0.077)	(0.513)	-	-	(0.495)	(0.054)	(0.104)
	2	-0.0010	-0.0011	-0.0015*	0.0011	0.0012	-0.0010	-0.0000	-0.0002
		(0.180)	(0.153)	(0.015)	(0.278)	(0.296)	(0.167)	(0.970)	(0.872)
	3	-0.0001	-0.0004	-0.0006*	0.0021*	0.0025*	0.0001	-0.0008	-0.0001
		(0.877)	(0.562)	(0.037)	(0.028)	(0.036)	(0.920)	(0.348)	(0.871)
	4	0.0001	-0.0006	-0.0003	0.0015	0.0004	0.0002	-0.0004	0.0000
		(0.749)	(0.068)	(0.069)	(0.053)	(0.544)	(0.736)	(0.357)	(0.873)
	5	0.0001	-0.0004	-0.0001	0.0013	-0.0006	0.0002	0.0000	0.0000
		(0.648)	(0.164)	(0.140)	(0.055)	(0.246)	(0.621)	(0.937)	(0.875)
	6	0.0000	-0.0003	-0.0001	0.0009	-0.0004	0.0001	0.0000	0.0000
		(0.755)	(0.163)	(0.227)	(0.076)	(0.323)	(0.845)	(0.983)	(0.877)
	7	0.0000	-0.0002	0.0000	0.0007	0.0000	0.0000	-0.0001	0.0000
		(0.629)	(0.242)	(0.313)	(0.097)	(0.971)	(0.884)	(0.754)	(0.879)

* Significance at the 5% level

We observe a significant positive relationship from positive tweets to returns the following day. From emotional tweets to returns, there is a significant negative relationship at lag 3 and 4. There is also a significant positive relationship from negative tweets and returns to positive tweets at lag 3.

Figure 14 below visualizes all significant relationships in Table 13.

**Figure 14 - Impulse Response Function Relationships**
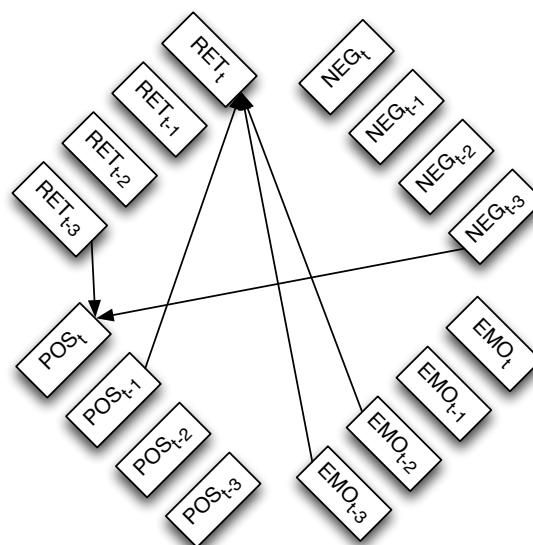
To gain a better understanding of the reactions we found in Table 13, we graph the effects over time. In Figure 15 below we can see that the response of an increase in the number of positive tweets is mixed, while the response to an increase in the number of negative tweets seems to be solely negative. However, only the effect from POS is significant, as is indicated when both red lines are at the same side of the x-axis. The change in sign in the RET to POS graph may be explained by an overreaction in the first period, followed by a correction in the second period. This could explain why the Granger causality test does not indicate a significant relationship from POS to RET, since it consider all lags in combination, and lag 2 abates some of the effect from lag 1.

Figure 15 - Impulse Response of RET to POS and NEG
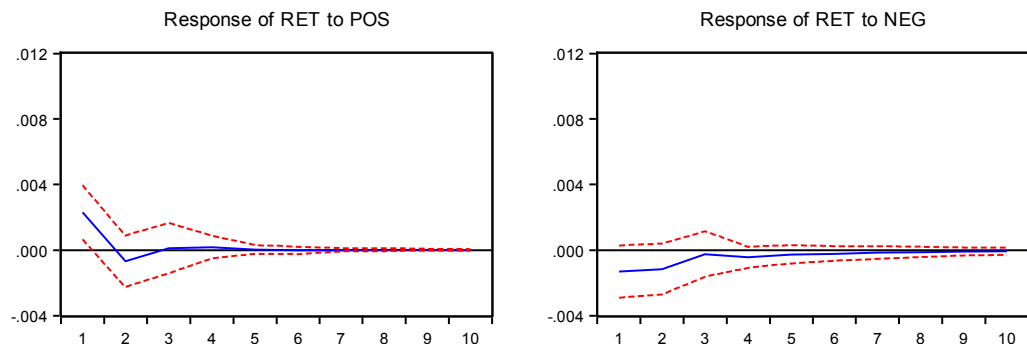
From Figure 16, we see that the overall effect an impulse in EMO has on RET is negative and peaks on day two, before diminishing over the next couple of days. The effect is significant at lag 2 and 3.

Figure 16 - Impulse Response of RET to EMO
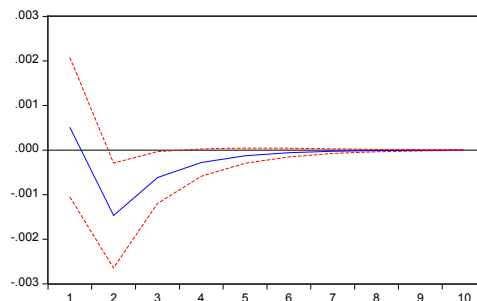
In Figure 17 below, we see that POS increases with a shock to both NEG and RET, both significant at lag 3. The first could be explained by negative tweets spurring discussion, which entails both more positive and negative tweets. The

fact that we observe an increase in the share of positive tweets after an increase in stock returns, but no effect on negative tweets, could be explained by investors bragging over their positive stock returns, while tending to keep quiet when they experience losses.

Figure 17 - Impulse Response of POS to NEG and RET
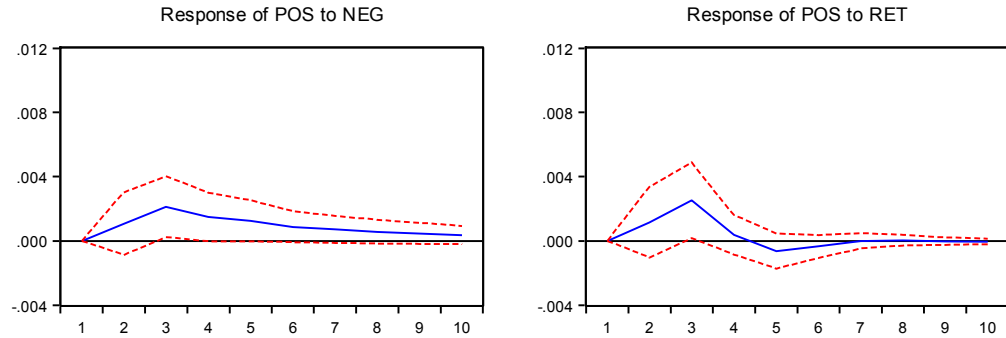
To better understand the longer term implications of the effects changes in our tweet variables have on stock returns, we also calculate the cumulative effect from each of the variables. In Table 14, the first row is identical to the corresponding row in Table 13. Beyond the first day, the only significant effect on RET is from NEG. After seven days, a standard deviation increase in the share of negative tweets leads to a cumulative reduction in stock returns of 0.44 percentage points.

Table 14 - Accumulated Impulse Responses of RET

This table summarizes the cumulative response in RET from a one standard deviation shock in EMO, POS and NEG. P-values are shown in parenthesis.

Dependent		RET		
Independent		POS	NEG	EMO
Lag	1	0.0018*	-0.0014	0.0005
		(0.025)	(0.077)	(0.513)
	2	0.0008	-0.0025*	-0.0010
		(0.476)	(0.023)	(0.323)
	3	0.0007	-0.0029*	-0.0016
		(0.58)	(0.017)	(0.161)
	4	0.0008	-0.0034*	-0.0019
		(0.546)	(0.014)	(0.13)
	5	0.0009	-0.0038*	-0.0020
		(0.505)	(0.017)	(0.12)
	6	0.0008	-0.0041*	-0.0020
		(0.522)	(0.021)	(0.115)
	7	0.0008	-0.0044*	-0.0021
		(0.553)	(0.025)	(0.116)

*Significance at the 5% level

The cumulative graphical representations of impulse responses in Figure 18 show that POS has a significant effect on RET at lag 1, which is corrected at lag 2 and then disappears. The effect from NEG to RET becomes significant at lag 2 and accumulates over time.

Figure 18 - Accumulated Impulse Response of RET to POS and NEG
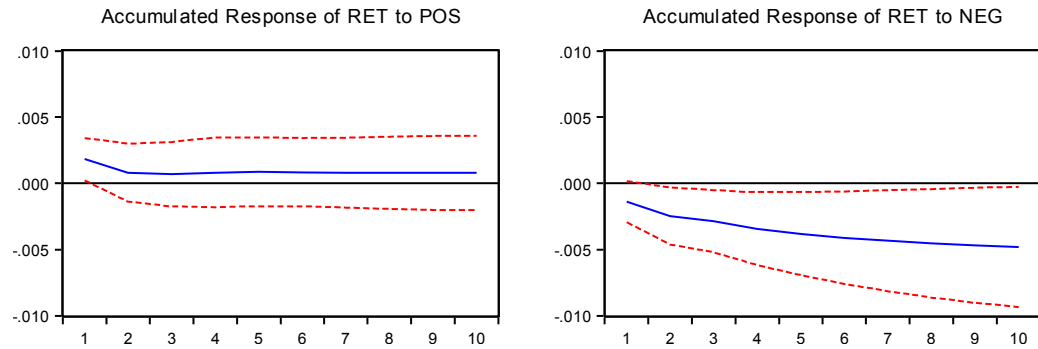
Figure 19 indicates that the long term effect on stock returns from a shock in the share of emotional tweets is negative. However, this effect is not significant.

Figure 19 - Accumulated Impulse Response of RET to EMO
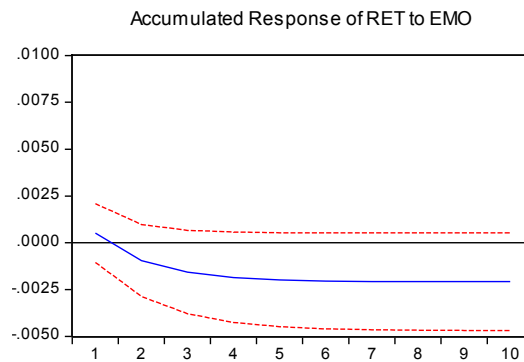
Variance Decomposition

Variance decompositions tell us how much of the forecast error variance (FEV) in RET is accounted for by innovations in the POS and NEG equations and how much is explained by own innovations in the RET equation (Lütkepohl 2006, 63). The variance decomposition for model 1 tells us that innovations in POS explains from 6.2 to 7.7 percent of the FEV in RET, while NEG explains from 3.6 to 6.8 percent of the FEV in RET, depending on the lags. Most of the FEV, however, comes from the own-shock in RET.

Table 15 - Variance Decomposition of POS, NEG and RET

This table shows how much of the forecast error variance in RET is explained by innovations in POS, NEG, and RET. The Cholesky ordering is POS NEG RET.

Period	S.E.	POS	NEG	RET
1	0.007	6.152	3.620	90.228
2	0.007	7.801	5.562	86.637
3	0.008	7.718	5.757	86.525
4	0.008	7.697	6.295	86.008
5	0.008	7.684	6.514	85.803
6	0.008	7.673	6.668	85.660
7	0.008	7.668	6.753	85.579

The variance decomposition for model 2 tells us that EMO explains from 0.5 to 5.4 percent of the FEV in RET.

Table 16 - Variance Decomposition of EMO and RET

This table shows how much of the forecast error variance in RET is explained by innovations in EMO and RET. The Cholesky ordering is EMO RET.

Period	S.E.	EMO	RET
1	0.012	0.517	99.483
2	0.013	4.548	95.452
3	0.013	5.226	94.774
4	0.013	5.366	94.634
5	0.013	5.394	94.606
6	0.007	5.400	94.600
7	0.007	5.401	94.599

The variance decompositions have shown that each of the sentiment variables explain a relatively small amount of the forecast error variance in RET.

Summary of Results

Our findings have the following implications for our hypotheses:

$H1_0$:   ~~Positive sentiment does not predict stock index returns~~

$H1_A$:   Positive sentiment predicts stock index returns

Although the VAR coefficients and the Granger causality test indicate no effect from the percentage of positive tweets to stock returns, the impulse response function table shows a significant positive reaction in returns one day after a positive shock of one standard deviation in the percentage of positive tweets. Thus

we reject the null hypothesis that positive sentiment does not predict stock index returns.

H2$_0$:   Negative sentiment does not predict stock index returns

H2$_A$:   Negative sentiment predicts stock index returns

The cumulative IRF indicates that an increase in the share of negative tweets has a long term effect on stock returns. We therefore reject the null hypothesis that negative sentiment does not predict stock index returns.

H3$_0$:   Emotional tweets do not predict stock index returns

H3$_A$:   Emotional tweets predict stock index returns

The strongest results are for the EMO variable. We reject the null hypothesis that the percentage of emotional tweets does not predict stock index returns both from the Granger causality test and from the impulse response function.

In addition we found that an increased share of negative tweets and an increase in the returns of the S&P 500 predict an increased share of positive tweets.

Conclusion

In this thesis our main objective was to test whether the sentiment of tweets can be used to predict stock returns. We obtained a unique data set of all tweets containing a stock ticker over a four-month period, and generated variables representing the sentiment contained in these tweets. We then created two VAR models where we estimated the relationships between these variables and the daily returns of the S&P 500 over the same period.

Our results indicate that a one standard deviation increase in the percentage of positive tweets leads to a 0.18 percentage point increase in stock returns the following day, while an increase in the percentage of emotional tweets predicts a reduction in stock returns two and three days later. Higher stock returns and a larger percentage of negative tweets both induce an increase in the percentage of positive tweets after three days. The only significant long-term effect we find is that a one standard deviation increase in the share of negative tweets leads to a 0.44 percentage point cumulative reduction in stock returns after seven days. From these results we infer that the markets react positively to positive sentiment, and negatively to negative sentiment and a higher level of emotions. The latter is in line with the findings of Bollen, Mao, and Zeng (2011) and Zhang, Fuehres, and Gloor (2011), who found a reduction in stock returns following increased levels of emotional sentiment. The magnitudes of our results are small but significant, and could thus be used to improve existing prediction models. Our results add to the evidence of the predictive powers of Twitter, which implies that market participants should pay attention to the information spread on Twitter, and may be able to profit from it.

One limitation of our study is that the normality of our models can be questioned. Model 1 did not pass the Jarque-Bera test of normality, but inspection of the residuals in a Q-Q plot revealed no serious deviation from normality. Non-normality should not increase the risk of making a "Type I" error, which is our main concern. Nevertheless, the confidence intervals, and thus p-values, may not be entirely correct.

The short time period of our data set is obviously limiting the power of our results. A longer data set could also help mitigate the problem of non-normality.

The Granger causality test, IRF and cumulative IRF gave very different results, which raises questions of the robustness of our results. While the Granger causality test indicated that only EMO could predict RET, the accumulated IRF indicated that only a change in NEG had a significant effect on RET in the long term. As with all prediction models, the results may be more or less significant if tested out of sample.

There are several possible and highly interesting avenues for future research in this field. An interesting option would be to verify the predictive models we have found by creating and simulating trading strategies and investigate their profitability. Developing custom sentiment classifiers accounting for different levels of sentiment strength and financial jargon might yield better predictions. In addition, including metadata such as the number of followers, enables weighting of content producers based on their influence and reach. Obtaining a longer and company-specific data set, scholars can perform analysis on the company level. Such analysis may provide insights into whether or not companies may benefit from actively using Twitter. Finally, we propose analysis on the intraday level. Because tweets are recorded with a granularity of one second, intraday level analysis could reveal how quickly the market reacts to tweets, allowing for a potentially very detailed examination of the relationships between Twitter and stock returns.

References

Antweiler, Werner, and Murray Z. Frank. 2004. "Is all that talk just noise? The information content of internet stock message boards." *The Journal of Finance,* 59 (3): 1259-1294.

Asur, Sitaram, and Bernardo A. Huberman. 2010. "Predicting the Future with Social Media". Hewlett Packard Social Computing Research. Accessed 26.09.12. http://arxiv.org/abs/1003.5699.

Baker, M, and J Wurgler. 2006. "Investor Sentiment and the Cross-Section of Stock Returns." *The Journal of Finance,* 61 (4): 1645-1680.

Bloomberg. 2013a. "Bloomberg Game Changers: Twitter." Accessed 19.06.2013. http://www.bloomberg.com/video/67515946-bloomberg-game-changers-twitter.html.

———. 2013b. "Bloomberg Integrates Live Twitter Feeds With Financial Platform." Accessed 19.06.2013. http://www.bloomberg.com/pressroom/bloomberg-integrates-live-twitter-feeds-with-financial-platform/.

Bodie, Zvi, Alex Kane, and Alan J Marcus. 2011. *Investments*. 9th ed. New York: McGraw-Hill Irwin.

Bollen, J, H Mao, and X Zeng. 2011. "Twitter mood predicts the stock market." *Journal of Computational Science,* 2 (1): 1-8.

Brooks, Chris. 2008. *Introductory Econometrics for Finance*. 2nd ed. New York: Cambridge University Press.

Bruns, Axel. 2012. "Ad Hoc innovation by users of social networks: The case of Twitter." *ZSI Discussion Paper,* 16 (2012): 1-13.

Coval, Joshua D, and Tyler Shumway. 2001. "Is sound just noise?" *The Journal of Finance,* 56 (5): 1887-1910.

DataSift. 2013. "Partners of DataSift." Accessed 15.08.13. http://powered.datasift.com/partners/lexalytics.

Dhar, V, and E Chang. 2007. "Does chatter matter? The impact of user-generated content on music sales". NYU Working Paper No. 2451/23783. Accessed 08.01.13. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1281347.

Edmans, Alex, Diego Garcia, and Øyvind Norli. 2007. "Sports sentiment and stock returns." *The Journal of Finance,* 62 (4): 1967-1998.

Fraser, Matthew, and Soumitra Dutta. 2008. "Obama's win means future elections must be fought online." Accessed 05.01.13. http://www.theguardian.com/technology/2008/nov/07/barackobama-uselections2008.

Gannes, Liz. 2011. "With Google Gone (For Now), Twitter Tries to Come to Terms With Microsoft's Bing." Accessed 10.06.13. http://allthingsd.com/20110715/with-google-gone-for-now-twitter-tries-to-come-to-terms-with-microsofts-bing/.

Ghiassi, M., J. Skinner, and D. Zimbra. 2013. "Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network." *Expert Systems with Applications,* 40 (16): 6266-6282.

Go, Alec, Richa Bhayani, and Lei Huang. 2009. "Twitter Sentiment Classification using Distant Supervision". CS224N Project Report. Accessed 12.06.13. http://cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf.

Hughes, Amanda Lee, and Leysia Palen. 2009. "Twitter adoption and use in mass convergence and emergency events." *International Journal of Emergency Management,* 6 (3): 248-260.

IBM. 2012. "IBM100 - Deep Blue." Accessed 30.04.13. http://www-03.ibm.com/ibm/history/ibm100/us/en/icons/deepblue/.

Jubbega, Annika. 2011. "Twitter as driver of stock price". Master Thesis, BI Norwegian Business School. Accessed 02.06.12. http://brage.bibsys.no/bi/handle/URN:NBN:no-bibsys_brage_29313.

Kanellos, Michael. 2003. "Moore's Law to roll on for another decade." Accessed 26.09.12. http://news.cnet.com/2100-1001-984051.html.

Kaplan, Andreas M., and Michael Haenlein. 2010. "Users of the world, unite! The challenges and opportunities of Social Media." *Business Horizons,* 53 (1): 59-68.

Kelly, Kate. 2011. "Hedge Fund to Predict Markets Using Twitter." Accessed 22.04.12. http://www.cnbc.com/id/41948279.

Kisling, Whitney, Eric Lam, and Nina Mehta. 2013. "Human Beats Machine This Time as Fake Report Roils Stocks." Accessed 24.05.13. http://www.bloomberg.com/news/2013-04-24/human-beats-machine-this-time-as-fake-report-roils-stocks.html.

Lane, Kin. 2012. "The Twitter Firehose." Accessed 28.09.12. http://apivoice.com/2012/07/12/the-twitter-firehose/.

Lütkepohl, Helmut. 2006. *New Introduction to Multiple Time Series Analysis*. Berlin: Springer.

Madrigal, Alexis C. 2013. "The Hashtag Is About to Roll Out to a Billion People, and This One Guy Invented It." Accessed 25.06.2013. http://www.theatlantic.com/technology/archive/2013/06/the-hashtag-is-about-to-roll-out-to-a-billion-people-and-this-one-guy-invented-it/276811/.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. "Thumbs up? Sentiment Classification using Machine Learning Techniques". Association for Computational Linguistics Conference. Accessed 15.06.13. http://www.cs.cornell.edu/home/llee/papers/sentiment.pdf.

PeopleBrowsr. 2012. "PeopleBrowsr Wins Temporary Restraining Order Compelling Twitter to Provide Firehose Access." Accessed 05.01.13. http://blog.peoplebrowsr.com/2012/11/peoplebrowsr-wins-temporary-restraining-order-compelling-twitter-to-provide-firehose-access/.

Ramos, Nicolas, and Nasif Islam. 2012. "Cognitive Science, Computer Science and Chess: Grandmaster Christiansen Visits C.U." Accessed 28.06.13. http://cornellsun.com/node/51143.

Redmore, Seth. 2012. "Machine Learning vs. Natural Language Processing, part 1." Accessed 15.08.13. http://www.lexalytics.com/lexablog/2012/machine-learning-vs-natural-language-processing-part-1.

Scannell, Kara. 2013. "Companies allowed to tweet #USearnings." Accessed 04.05.13. http://www.ft.com/intl/cms/s/0/99ccad94-9bd2-11e2-a820-00144feabdc0.html#axzz2Q3PwolB1.

Schmerken, Ivy. 2012. "Academic's Twitter Model Predicts Stock Returns." Accessed 22.04.12. http://www.advancedtrading.com/quant/academics-twitter-model-predicts-stock-r/232602958.

Sebastiani, Fabrizio. 2002. "Machine learning in automated text categorization." *ACM computing surveys (CSUR),* 34 (1): 1-47.

SEC. 2013a. *Report of Investigation Pursuant to Section 21(a) of the Securities Exchange Act of 1934: Netflix, Inc., and Reed Hastings*. Washington, D.C.: SEC.

―――. 2013b. "SEC Says Social Media OK for Company Announcements if Investors Are Alerted." 02.05.13. http://www.sec.gov/news/press/2013/2013-51.htm.

Sprenger, T, and I Welpe. 2010. "Tweets and Trades: The Information Content of Stock Microblogs". Working Paper. Accessed 22.04.12. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1702854.

Stenson, Ed. 2012. "Salience 5." Accessed 15.08.13. http://dev.datasift.com/blog/salience5.

Trejdify. 2012. "How could Paul Tudor Jones predict the Black Monday 1987 crash?" Accessed 26.08.13. http://blog.trejdify.com/2012/09/how-could-paul-tudor-jones-predict.html.

Twitter. 2013. "Audiences on Twitter." Accessed 29.05.13. http://business.twitter.com/audiences-twitter.

Whinston, Andrew, and Huaxia Rui. 2010. "Social Media as an Innovation -The Case of Twitter". Working Paper. Accessed 24.04.12. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1564205.

Wrenn, Eddie. 2012. "Twitter graphic shows how US produces 30% of all tweets (compared to censored China, with barely anyone to follow)." Accessed 05.01.13. http://www.dailymail.co.uk/sciencetech/article-2169204/Twitter-graphic-shows-US-produces-30-tweets.html.

Wysocki, Peter. 1998. "Cheap talk on the web: The determinants of postings on stock message boards". University of Michigan Business School Working Paper No. 98025. Accessed 18.06.13. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=160170.

Zeledon, Max. 2009. "StockTwits May Change How You Trade." Accessed 22.04.12. http://www.businessweek.com/technology/content/feb2009/tc20090210_875439.htm.

Zhang, Xue, Hauke Fuehres, and Peter A Gloor. 2011. "Predicting Stock Market Indicators Through Twitter "I hope it is not as bad as I fear"." *Procedia - Social and Behavioral Sciences,* 26: 55-62.

Jens-Otto Forbergskog
Christer Ryland Blom

Preliminary Thesis Report

Twitter and Stock Returns

Date of submission:
15.01.2013

Supervisor:
Prof. Øyvind Norli

BI Norwegian Business School

Oslo

Master of Science in Business and Economics, Major in Finance

GRA 19002 Thesis Preliminary Report

Table of Contents

Abstract

This preliminary thesis introduces our topic, motivation, and questions on which we will base our research. We provide a thorough review of the current literature on the topic before outlining our intended approach and methodology.

In our thesis, our main goal is to investigate whether linguistic analysis of tweets mentioning specific stock tickers can be used to predict the performance of those stocks. In particular we test for relationships with leading, current, and lagged abnormal returns, volatility and volume for the respective stocks. We obtain a longitudinal data set of tweets, from which we derive a set of explanatory variables using linguistic analysis. We then run regressions against historical stock return data from the same time period to test for such relationships.

Introduction

The use of social media as Twitter by investors is a recent phenomenon, and is therefore poorly covered by academic research. As we believe that social media will play an important role in the life of investors in the future, we feel it is an area worthy of academic study. We intend to contribute to the existing literature by providing an empirical test of stock market predictability using social media analysis.

There is obviously considerable monetary motivation in predicting the movements of the stock market, and many attempts have thus been made. However, few of the known predicting models are based on social media. Investors who are successfully predicting the market using such models, are unlikely willing to share their methods, as that would render them useless. Scholars, on the other hand, may be motivated to publish successful models to gain recognition and fame.

Automated trading based on algorithms analyzing real time market data from the stock exchanges is now quite common and well covered by research and media. Most such algorithms are however exclusively analyzing data provided by the stock exchanges themselves. Imagine instead a sophisticated algorithm able to access the entire pool of information known to man, including real time discourse in social media, and based on this make instantaneous, unbiased and rational investment decisions. We believe this could become the next generation of robot trading, which may replace not only traders, but also brokers and analysts. If this becomes a reality, computers will most definitively have an even more prominent role in the stock markets than they have today.

Up until recently it has not been possible to analyze big data (i.e. analysis of huge amounts of unstructured raw data) due to computational, storage and bandwidth constraints. The human brain has so far remained superior to computers in understanding the broader picture, but it has limited cognitive capacity and bounded rationality, and will eventually be surpassed.

The use of big data analysis by investors has already begun. Indeed, "some of the largest quant hedge funds, the likes of Renaissance Technologies, D.E. Shaw and others are said to be spending millions (if not billions) on building tools for analyzing unstructured data found on Twitter and Facebook. Big data companies like Thomson Reuters and Dow Jones are offering products and entire business units around interpreting sentiment analysis to produce trading signals" (Schmerken 2012). It is clearly big money in analyzing big data. Some of the most enthusiastic advocates of social media are of the opinion that it can be "construed as a form of collective wisdom" (Asur and Huberman 2010), and being able to tap into this collective wisdom should make for superior investment decisions.

In this preliminary thesis, we describe how we are going to use linguistic analysis of tweets and regressions to examine whether there are any significant relationships between stock specific tweets and stock performance, and if such relationships can be used to predict the movements of stocks. We begin by reviewing current academic literature on behavioral finance, market efficiency, big data, social media, and linguistic analysis. Finally, we introduce our hypotheses, data set and methodology.

Literature Review

Twitter is a global online social media service used by millions of individuals and organizations to exchange short messages of up to 140 characters. It has rapidly evolved to become a complete ecosystem and a powerful tool in several areas such as news, politics, health, and in our case, finance.

The growth of Twitter has been extraordinary. There are now officially over 140 million active Twitter users (Twitter 2012), and tweet volume has doubled in the last year, from 230 million daily tweets at the end of 2011 (Lane 2012) to more than 400 million tweets per day in November 2012 (PeopleBrowsr 2012).

The core of Twitter is called the *firehose*, which is the constant stream of tweets, with additional metadata embedded such as author, location and language. Even

though tweets in principle are public, access to the firehose is restricted, and has recently been tightened. This has lead to some angry partners such as PeopleBrowsr, who as a result sued Twitter. They won a restraining order in November 2012, forcing Twitter to continue providing firehose access (PeopleBrowsr 2012). While Microsoft seems to have paid the USD 30 Million Twitter reportedly demanded to retain firehose access for their search engine Bing, negotiations with Google failed and they are therefore no longer offering real time Twitter search (Gannes 2011). There are some select distributors, such as Gnip and Mediasift, who are allowed to resell their firehose access to third parties (Lane 2012). These distributors also keep complete records of all tweets, which can be accessed for historical data analysis.

There can be identified several different streams of research on Twitter. One stream focuses on understanding its usage and community structure. Another focus on the influence Twitter users have, in for example changing the outcome of an election, expose unethical behavior by companies, uncovering scandals, and affecting product sales. Our research belongs to another stream, which focuses on Twitter's prediction power and potential application to other areas.

Some of most relevant literature for our thesis is found in the field of behavioral finance. However, some of the studies on the stock market predictability using social media belong to the field of computational science. This is probably because such studies are computationally complex to conduct, and require specific skills in data manipulation.

The Predictability of Stock Markets

As mentioned before, it is extremely desirable to be able to predict the stock market, and a myriad of models have been developed for this purpose. Some of these models successfully predict returns for past data, but often fails in later attempts (Bodie, Kane, and Marcus 2011, 367). One possible explanation for this could be that the market adjusts for these new methods, through arbitrage, so that

they are no longer profitable. This discourages investors from sharing successful models.

Stock market analysis was one of the first applications of computers in economics. Maurice Kendall (1953) famously found that "he could identify no predictable patterns in stock prices. Prices seemed to evolve randomly" (Bodie, Kane, and Marcus 2011, 343). Such findings eventually led to the development of the efficient market hypothesis, which states that stocks already reflect all available information, making it impossible to predict their movement based on past data (Bodie, Kane, and Marcus 2011, 345).

While conventional financial theory usually assumes full rationality and efficiency, consensus in behavioral finance is that psychology and emotions are important factors in determining how investors behave. This implies that markets are not efficient, which is imperative for technical analysis to work. Modern behavioral finance theory suggests that humans are not rational machines, but emotional, rationally bounded, and subjective actors, who are influenced by other things than the cold facts (Bodie, Kane, and Marcus 2011, 382). Several studies find that psychology affects investor behavior. One example is a study that found "a significant market decline after soccer losses" (Edmans, Garcia, and Norli 2007), which shows that the mood of investors may influence the stock market.

High frequency algorithm trading is based on exploiting pricing errors and arbitrage opportunities in the market. It is thus actually founded on market inefficiencies, and inefficient markets are therefore a prerequisite for it to be profitable. As the number of robots engaged in the market increases, it becomes harder for them to be profitable. Big data analysis, on the other hand, exploits the inability of investors to consider all relevant information, and could be profitable even if the market is weak-form efficient. Unless insider information is leaked through social media, such analysis will not work if the market is semi-strong form efficient. Thus, our results will indicate a certain level of efficiency. Indeed, if we are able to predict the market, our study will add to the list of proof for market inefficiency.

Big Data Analysis

Our research is a form of big data analysis, one of the hottest concepts of recent time. Harvard Business Review had the spotlight on big data in their October 2012 issue, naming it "The Management Revolution". We take an investor rather than manager perspective, but the concept is the same; "data-driven decisions are better decisions … Using big data enables managers to decide on the basis of evidence rather than intuition" (McAfee and Brynjolfsson 2012).

We mentioned in the introduction that investors already are using big data analysis to make investment decisions. However, this data has to come from somewhere. One of the pioneers in using Twitter for stock related chatter is StockTwits. It has been said that "StockTwits is the modern version of traders shouting in the pits" (Zeledon 2009). It is thus the traders themselves who are the data source in this case. The discourse about stocks can in itself be valuable information, as it indicates investor interest.

Compared to the forums that predate modern social media for user-generated content (UGC) online, Twitter has powerful self-moderating aspects that the users appreciate. "By letting each individual user decide whom to follow, the content is moderated automatically: 'Underperformers will be ignored, and rightly so— trading is a zero-sum game and bad advice is a waste of time and money. That's precisely what validates apps like StockTwits.'" (Zeledon 2009). In addition, trusted sources such as major news agencies, governments and companies have verified accounts so that users can know the information in tweets from these sources are backed by more than just a nickname. A recent trend we observe is that the companies themselves are beginning to publish price-sensitive information in announcements on Twitter. Of course, such information is usually published through several channels simultaneously, making it impossible to isolate the effect of the tweet.

Although many news sites allow comments on their articles, there is demand by investment professionals for a common and independent place to discuss the news. "Traders and investors alike have come to view these platforms as trusted

filters that help them make more informed decisions because they can discuss and interpret the news with their peers" (Zeledon 2009). Twitter is one of the platforms that satisfy this demand.

All stocks are not equally appropriate for Twitter analysis. While unknown companies rarely are discussed on Twitter, the opposite problem arises with large companies, where the wide discussion and rigorous analyst coverage seems to make them too efficiently priced. Baker and Wurgler (2006) argue that small stocks, which are difficult to arbitrage, are more likely to be affected by emotion than highly liquid stocks. They hypothesize and find that "investor sentiment has larger effects on securities whose valuations are highly subjective and difficult to arbitrage". Small stocks are also more likely mispriced because they are often ignored by analysts: "Small stocks that receive relatively little coverage by Wall Street analysts may be less efficiently priced than large ones" (Bodie, Kane, and Marcus 2011, 346).

One of the primary factors enabling big data analysis is the rapid development of information technology. In particular, the increase in computational power has been exponential the last decades, with capacity remarkably closely following Moore's law, doubling every 18 months (Kanellos 2003). If this development continues, computers will eventually outperform humans in big data analysis, as it has done in other areas humans traditionally have had the edge, like checkers, chess, Scrabble and Jeopardy!. The first time a computer beat a top human being in chess was February 10, 1996 (IBM 2012). The last time a top human beat a top computer in chess was on November 21, 2005, and this will probably never happen again. Indeed, even modern mobile phones are now able to reach grandmaster level in chess (Crowther 2012).

Automated trading is not without problems. People have attempted and succeeded in tricking high frequency algorithm trading, including a famous case in Norway where two day-traders successfully profited from deceiving the robot Timber Hill. The supreme court of Norway found them not guilty of market manipulation (Norges Høyesterett 2012). This ruling affirmed that every market participant is

responsible for their own actions – fooling stupid robots should not be illegal. As long as there is money to gain, there will be incentives to trick big data algorithms as well. For example, this could be done by distributing false rumors, as happened during super-storm Irene, where several false reports on the development of the storm appeared on Twitter. Some of these were even quoted on CNN. However, the self-moderating effect of Twitter quickly buried these false reports when people discovered they were untrue (Ingram 2012).

The Forecasting Power of Twitter

Our focus is on finding out whether Twitter can be used to predict stock prices. There are a few studies on the predicting power of Twitter. Most, however, are not finance related. Nevertheless, we include some of them due to the rarity of such studies.

A paper by Asur and Hubberman (2010) demonstrates how Twitter data can be used to forecast box-office revenues for movies. They found that "a simple model built from the rate at which tweets are created about particular topics can outperform market-based predictors". Another study found that "the volume of blog posts about an album is positively correlated with future sales" (Dhar and Chang 2007). It is unclear to what degree social media are serving as a proxy for existing market interest or actually in itself leads to increased publicity and sales. The main point, however, is that both studies are successfully able to use social media to forecast sales.

One highly relevant study "investigate whether measurements of collective mood states derived from large-scale Twitter feeds are correlated to the value of the Dow Jones Industrial Average (DJIA) over time" (Bollen, Mao, and Zeng 2011). They used the popular mood-tracking tool OptionFinder as well as their own tool, GPOMS, and found predicting power for some of the public mood dimensions. OptionFinder did not prove to be particularly effective in predicting the DJIA, but the GPOMS dimension "calmness" was a good predictor.

A study by Sprenger and Welpe (2010) is also closely related to ours. They found "the sentiment (i.e., bullishness) of tweets to be associated with abnormal stock returns and message volume to predict next-day trading volume". This study differs from ours in that it focuses on the market level rather than considering individual stocks.

In her master thesis, Jubbega (2011) found that brand sentiment tweets had an effect on the stock price for 5 of 10 companies. She found that investor reactions grow over time, peaking after 2 to 4 days, then declining 1 to 6 days after the peak. Our study differs in that it takes a finance rather than marketing perspective. Also, instead of using mentions of brands, we use mentions of the stock ticker. This gives us a data set of tweets specifically concerning the stock, instead of general discourse about the company or its brands.

Linguistic Analysis for Sentiment Determination

A common approach to testing the predictable power of Twitter is tracking the sentiment of tweets. The sentiment of tweets can be viewed as a proxy for the general mood in the market, which as shown above can affect stock prices.

In determining the sentiment, the factors that induce people to use positive and negative wording have to be considered. One study found that the specific words people use in tweets are not only related to their opinion of whether to buy or sell a certain stock, but also dependent on the general mood: "people start using more emotional words such as hope, fear and worry in times of economic uncertainty, independent of whether they have a positive or negative context" (Zhang, Fuehres, and Gloor 2011). Thus, volatile periods could be predicted by measuring the amount of emotional words.

Methodology

We intend to use a Twitter database service like Gnit to obtain a data set of tweets within our specified range and criteria, which we process using the programming language Perl to generate a series of variables for our study. We will obtain stock

return data for the same period from Yahoo! Finance. Then we will run a series of multivariate regressions to determine whether the variables have any leading, current or lagging relationships with stock returns, volatility or volume. We use stock price, volume and volatility as the dependent variables because they are of high interest to investors, and are easy to access and measure.

Hypotheses

We present the following hypotheses:

$H1_0$:   Positive sentiment in a specific day is unrelated to the stock return the preceding, current, or following day.

$H1_A$:   Positive sentiment in a specific day is positively related to the stock return the preceding, current, or following day.

$H2_0$:   Negative sentiment in a specific day is unrelated to the stock return the preceding, current, or following day.

$H2_A$:   Negative sentiment in a specific day is negatively related to the stock return the preceding, current, or following day.

$H3_0$:   The relative volume of tweets about a company in a specific day is unrelated to stock returns, stock volatility and volume the preceding, current, or following day.

$H3_A$:   The relative volume of tweets about a company in a specific day is unrelated to stock returns, but is positively related to stock volatility and volume the preceding, current, or following day.

Data

We have chosen to focus exclusively on the US market, mainly because it is currently Twitter's largest market by far. In fact over 30% of tweets are originating from within the US (Wrenn 2012). In particular, discourse specific to stocks seems to be mostly a US phenomenon for now.

We aim to analyze at least 20 stocks in at least two different size groups. We want to include both large and small companies in our study to uncover any differences in predictability.

Because Twitter does not provide free access to the firehose of all tweets, we have to rely on a third party commercial service for our tweet data. We intend to use a service like Gnit, which maintains a complete historic database of all tweets.

We will use so-called "cashtags", created by putting the dollar sign before the ticker (e.g. $AAPL), as search strings to obtain stock relevant tweets. Cashtags was initially invented by stocktwits.com, but has later been officially adopted by Twitter (Bohn 2012). Using only tweets that contains cashtags will not give us all tweets relevant to a given company, but we believe this is the best approach to obtain a useable data set for our study.

Daily historical data on return and volume for specific stocks are obtained from Yahoo! Finance. We use this data to calculate volatility.

Analysis

We will rely on standard or custom algorithms for determining the sentiment of tweets. There are several tools available to determine the sentiment of tweets, such as OpinionFinder. Most such tools are based on the simple concept of counting positive and negative laden words, while some utilize more sophisticated algorithms. To test our hypotheses, we intend to analyze several aspects of the tweets:

- The number of tweets in a specific day.
- The number of tweets in a specific day relative to the monthly average.
- The number of negative tweets versus the number of positive tweets in a specific day, on a scale from -1 to 1, where -1 means 100% negative tweets and 1 means 100% positive tweets.

Regressions

We run a series of multivariate regressions on the variables we generated to test if our hypotheses are supported by data.

Our Twitter variables are as follows:

TVOLU   Tweet volume

RTVOLU   Relative tweet volume

POSNEG   Positive versus negative sentiment

Our stock variables are as follows:

RET   Stock return

VOLA   Stock volatility

VOLU   Stock volume

We have arrived at the following initial regression models:

RET   = TVOLU RTVOLU POSNEG

VOLA = TVOLU RTVOLU POSNEG

VOLU = TVOLU RTVOLU POSNEG

Our conclusions will be based on the results of these regressions. Depending on the results, we may run additional statistical tests.

Progression Plan

We have outlined a plan for our thesis work in the table below, including the expected timeframe for each part.

Thesis Work	Timeframe
Obtain data set (tweets and stock data)	January-February 2013
Process and analyze data using Perl	February-March 2013
Define models, run regressions, statistical tests	March-April 2013
Interpret and comment on results, draw conclusions	April-May 2013
Hand in for feedback, review, refine, proofread	May-August 2013
Deadline	September 1, 2013

References

Asur, S, and B A Huberman. 2010. "Predicting the Future with Social Media."
    Hewlett Packard Social Computing Research.

Baker, M, and J Wurgler. 2006. "Investor Sentiment and the Cross-Section of
    Stock Returns." *The Journal of Finance* 61 (4).

Bodie, Zvi, Alex Kane, and Alan J Marcus. 2011. *Investments*. 9 ed. McGraw-Hill
    Irwin.

Bohn, Dieter. 2012. "Twitter Makes Stock Symbol $ 'Cashtag' Links Official,
    Following # and @ | the Verge." *Theverge.com*.
    http://www.theverge.com/2012/7/30/3205284/twitter-stock-ticker-cashtag-
    links-official.

Bollen, J, H Mao, and X Zeng. 2011. "Twitter Mood Predicts the Stock Market."
    Journal of Computational Science 2 (1).

Crowther, Mark. 2012. "The Week in Chess 771." *Chess.Co.Uk*.
    http://www.chess.co.uk/twic/twic771.html#13.

Dhar, V, and E Chang. 2007. "Does Chatter Matter? the Impact of User-
    Generated Content on Music Sales."

Edmans, A, D Garcia, and Ø Norli. 2007. "Sports Sentiment and Stock Returns."
    The Journal of Finance 62 (4).

Gannes, Liz. 2011. "With Google Gone (for Now), Twitter Tries to Come to
    Terms with Microsoft's Bing." *AllThingsD.com*.
    http://allthingsd.com/20110715/with-google-gone-for-now-twitter-tries-to-
    come-to-terms-with-microsofts-bing/.

IBM. 2012. "IBM100 - Deep Blue." *Www-03.Ibm.com*. http://www-
    03.ibm.com/ibm/history/ibm100/us/en/icons/deepblue/.

Ingram, Mathew. 2012. "Hurricane Sandy and Twitter as a Self-Cleaning Oven
    for News." *Gigaom.com*. http://gigaom.com/2012/10/30/hurricane-sandy-and-
    twitter-as-a-self-cleaning-oven-for-news/.

Jubbega, Annika. 2011. "Twitter as Driver of Stock Price" (September 1).

Kanellos, Michael. 2003. "Moore's Law to Roll on for Another Decade."
    News.Cnet.com. http://news.cnet.com/2100-1001-984051.html.

Lane, Kin. 2012. "The Twitter Firehose." *Apivoice.com*.
    http://apivoice.com/2012/07/12/the-twitter-firehose/.

McAfee, A, and E Brynjolfsson. 2012. "Big Data: the Management Revolution."
    Harvard Business Review (October).

Norges Høyesterett, ed. 2012. *Hr-2012-00919-A* Norges Høyesterett.

PeopleBrowsr. 2012. "PeopleBrowsr Wins Temporary Restraining Order Compelling Twitter to Provide Firehose Access." *Blog.Peoplebrowsr.com*. http://blog.peoplebrowsr.com/2012/11/peoplebrowsr-wins-temporary-restraining-order-compelling-twitter-to-provide-firehose-access/.

Schmerken, Ivy. 2012. "Academic's Twitter Model Predicts Stock Returns - Advanced Trading." *Advancedtrading.com*. http://www.advancedtrading.com/quant/academics-twitter-model-predicts-stock-r/232602958.

Sprenger, T, and I Welpe. 2010. "Tweets and Trades: the Information Content of Stock Microblogs." *Working Paper*.

Twitter. 2012. "Twitter Turns Six." *Blog.Twitter.com*. http://blog.twitter.com/2012/03/twitter-turns-six.html.

Wrenn, Eddie. 2012. "Twitter Graphic Shows How US Produces 30% of All Tweets | Mail Online." *Dailymail.Co.Uk*. http://www.dailymail.co.uk/sciencetech/article-2169204/Twitter-graphic-shows-US-produces-30-tweets.html.

Zeledon, Max. 2009. "StockTwits May Change How You Trade." *Businessweek.com*. http://www.businessweek.com/technology/content/feb2009/tc20090210_8754 39.htm.

Zhang, Xue, Hauke Fuehres, and Peter A Gloor. 2011. "Predicting Stock Market Indicators Through Twitter 'I Hope It Is Not as Bad as I Fear'." *Procedia - Social and Behavioral Sciences* 26 (January).