KU LEUVEN

FACULTEIT
INGENIEURSWETENSCHAPPEN

Predicting the stock market by using sentiment analysis on Twitter

Jeroen Ruytings

Thesis voorgedragen tot het behalen
van de graad van Master of Science
in de ingenieurswetenschappen:
computerwetenschappen,
hoofdspecialisatie Artificiële
intelligentie

Promotoren:

Prof. dr. M.-F. Moens
Prof. dr. ir. M. Van Barel

Assessoren:

Dr. ir. J. Ramon
Dr. O. Kolomiyets

Begeleider:

Dr. J. C. Gomez

Academiejaar 2013 – 2014

Preface

There are a number of persons without whom this master thesis might never be written, so I would like to say a few words of gratitude towards them.

In the first place I would like to thank my promoters, professor dr. Sien Moens and professor dr. ir. Marc Van Barel. They gave me the necessary support and advice to do this research. They sent me interesting papers on the topic and gave me useful feedback during the different meetings.

I would like to thank my daily supervisor dr. Juan Carlos Gomez. He guided me from the beginning until the end and was a great support to me. During our weekly meetings, he gave not only useful feedback on my ideas, he also shared some interesting new ways of looking to the problem.

My gratitude goes also to my parents. They made it possible to follow this studies in computer science. They always encouraged me, not only during this master thesis, but during my full academic career.

I would like to express my gratitude to Richard Vandergrift of the company RC Research, who provided me the necessary stock data to perform my different experiments. Without that data it would be impossible to do the experiments.

I would like to end by thanking my friends and family, who gave me the necessary distraction when I needed it and who motivated me through difficult moments.

Jeroen Ruytings

Contents

Abstract

Stock market investors would have a great help if a system would tell them if the stock market will go up or down in the future. It would give them the opportunity to invest smartly and gain much more profit. However very desired, such a system is not easy to make. The stock market is characterized by its complexity, with economic, political and psychological factors influencing the stock movements. This master thesis tries to find out if Twitter is also part of those factors that have an influence on the stock market.

The thesis starts with a literature study on previous work. Multiple researchers have studied the problem, using news articles, financial reports or microblogging posts to predict the stock market. The researchers do not agree on the best technique to predict the stock market, but they all claim to have promising results, showing correlations between their used source and the stock market fluctuations. The reality shows that their systems are rarely making profit, maybe due to the fact that the stock market is very complex and does not depend on a single factor.

Different experiments were done to get a better understanding of Twitter's role in stock market prediction. The experiments show that a window of influence of 10 minutes could exist, meaning a period for which tweets might have an impact on stock prices. From the experiments with the data used in this thesis, no total returning correlation between tweets and stock market in a whole day was found. Nevertheless, for some moments of the day the correlation between tweets and stock market becomes significant for a longer period of time. This might suggest that the influence of Twitter on the stock market varies during the different time periods (days, weeks or even years).

List of Figures

List of Tables

List of Abbreviations and Ticker Symbols

Abbreviations

API	application programming interface
AZFinText	Arizona Financial Text System
DCL	data control language
DDL	data definition language
DJIA	Dow Jones Industrial Average
DML	data manipulation language
EM	expectation-maximization
EOD	end-of-day
EMH	efficient market hypothesis
forex	foreign exchange market
GDP	gross domestic product
GICS	Global Industry Classification Standard
GPOMS	Google-Profile of Mood States
HSI	Hang Seng Index
IPO	initial public offering
LDA	latent Dirichlet allocation
LSA	latent semantic analysis
NASDAQ	National Association of Securities Dealers Automated Quotations
NLP	natural language processing
NYSE	New York Stock Exchange
POS	Part-of-speech
REST	Representational state transfer
ROC	receiver operating characteristic
S3VM	semi-supervised support vector machines
S&P 500	Standard & Poor's 500
SQL	structured query language
SVM	support vector machine

Ticker symbols

AAPL	Apple
GOOG	Google
MSFT	Microsoft
TWTR	Twitter
VIX	Chicago Board Options Exchange Market Volatility Index

Chapter 1

Introduction

1.1 Stock market prediction

Stock exchanges have fascinated people since their beginning. They provide to companies an opportunity to raise money, while they let investors dream of big profits. The stock market is characterized by its complexity, with economic, political and psychological factors influencing the stock movements. The size of the world stock market capitalization appeals to the imagination: the year 2012 closed with a total capitalization of 54.6 trillion dollar [42].

Many investors take their chance and invest in one or more stocks in the hope to sell them later at a higher price. This is not without risks: prices fluctuate constantly and might hence also go down. In the recent crisis, a lot of people lost money in this way. Nevertheless the stock market keeps attracting people who are looking for the big fortune.

Investors would have a great help if a system would tell them if the stock market will go up or down in the future. It would give them the opportunity to invest smartly and gain much more profit. It is also something that attracts researchers. Since the sixties researchers try to discover if it is theoretically possible to predict the stock market or not.

In 1997 the first research projects appeared that tried to forecast the stock market based on news articles. During the years that followed, more and more refinements on the first prediction system were introduced within the research area, helped by the increasing availability of textual sources on the World Wide Web. Soon also other textual sources were used for the research, for instance financial reports and microblogging posts.

Despite the attempts of researchers, nobody has found the holy grail of stock market prediction yet. Several projects obtain promising results, but they are rarely profitable. Most systems suffer from high transaction costs, although those transaction costs are often ignored in the papers of researchers.

1.2 Twitter

Since Twitter was put online in 2007, it has become a more and more important social network and online microblogging service. Nowadays, Twitter has become one of the most popular media to share opinions and thoughts with the world. Twitter has currently 255 million monthly active users. They post each day 500 million tweets [56]. When an important event occurs somewhere in the world, it becomes immediately a trending topic on Twitter.

The popularity of Twitter provides an opportunity for stock market investors. If they could find useful information in the overwhelming amount of Twitter data, they could gain an important advantage over people who do not have that information. Processing information faster than your opponents is essential for investors.

Twitter has also some disadvantages. People post almost everything on Twitter. A lot of tweets are noise and not useful for stock market prediction. This is a big challenge in the domain. Methods focussing on financial tweets, can filter those tweets out of a stream of tweets by filtering on the dollar sign ($) followed by the ticker of a company. This is a convention between Twitter users to talk about financial news of companies.

As said before, there are many factors influencing the stock market. It is not clear if tweets are one of those factors. It is also possible that tweets have a temporal effect on the stock market and not a continuously effect. That means, they are important only during a limited amount of time (and perhaps only for very important events).

1.3 Sentiment analysis

One possibility to increase the performance of prediction systems is investing more research to the understanding of textual sources. If systems can make a better distinction between positive and negative news or between useful or not useful, they will also be able to predict better the reaction of the stock market.

Sentiment analysis is the task of detecting the polarity of a text. The task has typically three parts: finding proper data, defining the sentiment features and executing an algorithm to perform the actual sentiment analysis task. Literature has not found a consensus on the best sentiment features nor the best algorithm. This provides opportunities for further research.

A deeper insight in the sentiment of texts allows researchers to study better the reaction of the stock market. As stated before, the stock market is a very complex system that is influenced by many factors. It is unclear if the sentiment of tweets is one of those influencing factors.

1.4 Goal of this master thesis

1.4.1 Literature review

The first goal of this master thesis is trying to get a better understanding in the domain of stock market prediction by studying the literature.

It is impossible to predict the stock market without having some basic knowledge about the economic principles that hide behind it. Chapter 2 provides this necessary background knowledge. It starts with explaining the role of financial markets in our economic system and then zooms in on the stock market, giving insights in the trade of stocks. The chapter continues with a theoretical overview on predicting the stock market and ends with an explanation on the two different opinions in how to put this theory into practice. The first group of analysts believe that a fundamental analysis is the right way to go, the second group uses a technical analysis.

This master thesis uses sentiment analysis of tweets to predict the stock market. Sentiment analysis tasks typically consist of three parts. Chapter 3 explains those three parts. The chapter starts with a motivation why a good data collection is essential to perform the task. It continues by explaining the different feature categories that can help to determine the sentiment. The chapter ends with an overview of the different algorithms that can be used to perform the sentiment analysis task and that can help to determine the sentiment of a text.

After explaining the necessary background knowledge on the economic principles and sentiment analysis, the thesis continues with a chronological overview of existing systems that try to predict the stock market based on textual data sources. Chapter 4 is divided in three parts, each clustering different approaches that are based on the same source of textual data. The chapter starts with approaches that use news articles, as this was the first data source that was used in the domain of stock market prediction. The chapter then explains how projects moved from using news articles to other data sources, like financial reports. The chapter ends with an overview on more recent papers using Twitter as its data source.

1.4.2 Sentiment analysis system

The second part of this master thesis involves the design of a sentiment analysis system. The target of this system is to get a better understanding of the role of Twitter as an influencing factor on the stock market.

All sentiment analysis tasks start with a data set. The data set that is used in this master thesis is described in chapter 5. The chapter explains first how the data is stored and logically structured in the database. It continues by describing the two different kinds of data that were collected: tweets and stock market data. Two Twitter data sets were found: one with labelled tweets, useful for performing supervised learning, and one with tweets containing financial news. For both data sets fine-grained stock market data of the same time period was necessary to study the link between tweets and the stock market. Huge amounts of tweets are appearing on any moment in time and their influence on the stock market might hence be more limited in time than news articles. Fine-grained data allows to find out of this hypothesis is true. Chapter 5 explains were the stock market data was found.

Chapter 6 explains the different experiments that were done to get a better understanding of Twitter's role in stock market prediction. The chapter starts by explaining the preprocessing phase. This phase removes some typical curiosities of tweets that don't contribute to the sentiment of the message. The preprocessing is followed by the experiments itself. The first set of experiments follows a clustering approach, where the sentiment is determined by using a clustering algorithm. Then topic models are used to get a better insight in the topics that appear in the positive, negative and neutral tweets and the words that are used within those topics. The last set of experiments uses the Stanford sentiment treebank, which can determine the sentiment of a sentence based on its structure. The tool is useful because it does not need manually labelled data.

Finally, chapter 7 recapitulates the two parts of this master thesis: the literature study and the own experiments. It resumes the most important results and it gives suggestions for future work.

Chapter 2

Economic principles

Before trying to better understand the relation between tweets and stock shares, it is important to know some of the basic ideas and principles that form the foundation of the stock market.

This chapter starts with an introductory section 2.1 on financial markets. This section describes the economic importance of financial markets and explains the different types of financial markets. The next section, section 2.2, focusses on the stock market, the research topic of this thesis. It gives information about shares, the trade of them and explains how indices are composed. Section 2.3 gives theoretical background about the feasibility of predicting the stock market. This results in two possible ways to predict the stock market, described in the following sections: section 2.4 talks about fundamental analysis, while section 2.5 give an insight in technical analysis. These two kinds of analysis try to determine which stocks to buy at which price. Fundamental analysts try to estimate the intrinsic value of a company and its shares based on economic indicators and use this as a basis for their prediction. Technical analysts use historical market data, such as volume and price, to forecast the stock market. The chapter ends with a conclusion.

2.1 Financial markets

Financial markets are markets where people and entities can trade financial assets such as equities, bonds, currencies and derivatives [35][64]. Such financial assets are often called *securities*. The prices of those securities are a reflection of the relation between supply and demand. Financial markets are typically defined by having transparent pricing, basic regulations on trading, costs and fees, and market forces determining the prices of the traded securities [26].

2.1.1 Importance

A healthy financial market is important to move funds from people who save to people who want to invest. Well-performing markets are an engine for economic growth. Greater economic growth has some advantages for daily life: more personal wealth, more jobs, better public services and more investments from which everybody can benefit.

2.1.2 Types

There are different ways to classify financial markets: you could divide them for instance into general and specialized markets, or into primary and secondary markets. The way they will be presented here is by what is traded in the market [26].

Capital market

On a capital market, individuals and instances can trade financial securities, typically on a long term. Organizations, institutions, companies and governments can sell securities to raise their funds. In general, there are two types of capital markets: stock markets and bond markets. The stock market allows investors to buy and sell equity securities, like stocks and shares, from publicly traded companies. The bond market allows investors to trade debt securities, like bonds. The bond market is often used by companies and governments to raise their income. This thesis focusses at stock markets, however here are also presented other types of markets to get a better understanding of the place of the stock market within financial markets.

Commodity market

At a commodity market, primary products rather than manufactured products are traded. Soft commodities are agricultural products like sugar or coffee, while hard commodities are mined, such as gold and oil.

Money market

The money market allows people to trade financial instruments with high liquidity on a short term, for one year or less. Due to the high liquidity, the money market is seen as a safe place to put money and the returns are significantly lower.

Derivatives market

Derivatives of other financial assets are traded at the derivative market. An example of derivatives are options, which give the holder the right to buy or sell securities at a specified price on a specified date.

Futures market

At futures markets people can trade in futures contracts. These contracts are contracts to buy specific quantities of financial products or commodities at a specified price on a specified time in the future.

Insurance market

The insurance market redistributes a wide variety of risks. An insurance company, or insurer, sells insurances to people who want to cover (part of) their risks. They pay a premium to the insurer, which will pay a compensation in case certain circumstances happen.

Foreign exchange market

The foreign exhange market or forex market is a place where currencies can be traded. The main participants are large international banks. The forex determines the value of the different currencies on a decentralized way.

2.2 Stock market

As introduced before, the stock market is a capital market on which equity securities are traded. The securities that are traded are shares of publicly traded companies. The price of those shares varies, reflecting demand and supply. If demand is higher than supply, prices will go up. If supply is higher than demand, prices will lower.

2.2.1 Shares

Companies can split up their capital into shares which can be sold at stock markets. When people buy a share, they become owners of a small part of the company and they will get a vote in the shareholder meeting to determine the strategy and direction of the company. All shareholders together are the owners of a company. Shareholders can earn money in two ways with their shares. The first way is when the company returns each year a part of their profit to the shareholders. This part is divided by the number of shares and the returned profit per share is called the *dividend*. A second way to earn money is to sell your share. The current market price will determine what you will get for your share when selling it. If the current market price is higher than the price on which you bought your share, you will earn money, in the other case you will lose money.

2.2.2 Trade

Initial public offering

The first time that shares of a particular company are sold to the general public is called the *initial public offering* (IPO). An investment banking firm helps to determine the correct price of the shares. After the IPO, the company transforms from a private company into a public company and shares can be traded according to the laws of supply and demand.

Ticker symbols

When a company goes to the stock market, it receives a symbol. This symbol is a way to recognize the company and is often an abbreviation of its name. The symbol helps investors to trade more easily in shares of the company. It can consist of letters, numbers of both, and it is unique for a given company, but can differ from market to market. The word "ticker" refers to the ticker tape, which was used as the earliest digital electronic communication medium to transfer stock price information over telegraph lines [66].

Brokers & dealers

Once launched on the stock market, shares can be traded according to supply and demand. Brokers act like agents during the purchase of securities. They match buyers and sellers. They are paid a brokerage commission during each transaction; this is why it is more efficient to trade in large quantities. Dealers link buyers and sellers by standing ready to buy and sell securities at given prices [35]. Dealers hold inventories of securities, this makes dealing a high-risk business.

Stock exchanges

A stock exchange is a market where buyers can meet sellers and where trades can happen on a centralized location. In this way the time between selling and buying shares is kept as short as possible. This is important because the smallest delay can have large economic consequences. Shares can be handled at different stock exchanges. Each stock exchange has opening hours and closure days.

The biggest stock exchange by market capitalization is the New York Stock Exchange (NYSE). It has a market capitalization of 17,950 trillion dollar [68]. This is the total value of shares of publicly traded companies at the stock exchange. In the beginning all transactions were the result of physical communication between people on a physical location. During the years this became more and more hybrid: some transactions are still physically done, but most transactions are virtually done.

The second biggest stock exchange is the National Association of Securities Dealers Automated Quotations (NASDAQ). This market is entirely virtual. All transactions are done with computers. It has a market capitalization of 6,085 trillion dollar[68]. High-technology companies are typically registered on NASDAQ.

The entry fee of NYSE is higher than that of the NASDAQ, but in return the NYSE provides more stability and more liquidity.

2.2.3 Indices

Stock market indices are weighted indices which are composed with different stock prices and which can be used to compare the market over the time. Most indices are composed of important shares, but there are also industry specific indices.

The oldest and probably most known index is the Dow Jones Industrial Average (DJIA). It consists of 30 major American companies. The index is a price-weighted index, which means that shares with higher prices contribute more to the index.

A second famous index is the NASDAQ Composite. It consists of all NASDAQ companies, which means that there are more than 3000 companies in it. These companies are both US and non-US because both can be on NASDAQ. The index is a market-value-weighted index, which means that companies with a higher market capitalization contribute more to the index.

Standard & Poor's 500 (S&P 500) is a third well-known index. It consists of 500 large American companies listed on NYSE or NASDAQ. Because of the large number of companies involved, it gives a good indication of how the American economy is performing. It is a market-value-weighted index, just like the NASDAQ Composite.

The 20 largest shares that are traded at the Brussels Stock Exchange are grouped in the BEL20-index.

2.3 Predicting the stock market

2.3.1 Price evolution

Stock prices are evolving in time according to the laws of supply and demand. But not only the market value of the company is important: there are other factors that can influence the stock prices. One important factor is the global evolution of the market. When the market is evolving in a positive direction, investors will be more confident to invest in shares and will push the stock prices even higher. In the other case, when the prices of the shares are evolving in a negative direction, investors will start selling their shares to avoid losing more money.

If people would be able to predict the stock prices, they would exactly know the right moment to buy or sell stocks in order to make as much profit as possible. However, this is not an easy task: investors are complex creatures and the number of external factors influencing the stock market is enormous.

2.3.2 Random walk hypothesis

The random walk hypothesis is a theory that states that the changes in the stock price have no memory [17]. The previous means that the stock history cannot be used to predict the future in a meaningful way. The future price level is no more predictable than a series of random numbers.

2013 Nobel Prize laureate Eugene F. Fama states in his work [17] that the random walk hypothesis is probably not entirely right. There might be small dependencies between successive price changes, however these dependancies may be so small that they can be considered as unimportant.

2.3.3 Efficient market hypothesis

The efficient market hypothesis (EMH) is a theory that is based on the random walk hypothesis. It assumes that all available information is already contained in the price of a share. The actual prices of securities reflect the effects of information based on events that have already occured and events that are expected to take place [17]. This implies that making large amounts of profit by doing predictions is impossible. Some people will indeed make more profit than others, but on the average this is normal distributed and the market is balanced.

The EMH can be divided in three forms: a strong, a semi-strong and a weak form. The distinction between those three forms is made on what people understand as "all information".

Weak form

In the weak form of the EMH only historical stock data is contained in the stock prices. Examples of this historical data are the historical evolution of the stock prices and the amounts of traded volumes. However, people cannot make profit systematically by trying to predict the stock prices, since the prices follow a random walk, as mentioned in section 2.3.2.

Semi-strong form

In the semi-weak form not only historical stock data, but also public available information is contained in the stock prices. News that appear will have an immediate impact on the stock price of a share. People cannot make profit by adapting their strategy based on new information, but they could make profit by using insider information.

Strong form

In the strong form all information is contained in the stock prices, whether it is public available or not. In many countries there exists an insider trading law, which states that people with insider information are not allowed to trade in shares of the company of which they have this information. This would imply that the strong form of EMH could not occur.

Usefulness

Studies do not agree of whether the EMH is applicable or not on the stock market. Some say that the weak form is applicable, others declare that the semi-strong form is applicable. It is clear that predictions based on new information, such as news articles or tweets, can only occur when EMH does not occur or at most in its weak form.

2.4 Fundamental analysis

The fundamental analysis of a company examines the financial health of a company, its strategy and its position in the market. The financial health of a firm can be determined from annual and quarterly reports or from audits by experts. The position of a company in the market depends on the market potential of its products and on the competitors in the market.

Based on this information, a fundamental analyst tries to determine the intrinsic value of a company and of its shares. If the intrinsic value of a company is higher than its current price on the stock market, then it is useful to buy shares because they will be profitable. If the intrinsic value is lower, then you need to sell your shares because the prognosis is that the stock price will go down.

Fundamental analysis can only occur when the EMH does not occur or if it occurs in its weak form. The semi-strong form and the strong form assume that the economic information where fundamental analysis is based on, is already contained in the stock prices.

A fundamental analyst can typically work in two ways: top-down or bottom-up, which will be explained in sections 2.4.1 and 2.4.2.

2.4.1 Top-down approach

The top-down approach starts at the macroeconomic level. Analysts look first to the global economy to see if the global climate is attractive for investors. When it is, they look into industries that are interesting. There can be industries that are doing it particularly well or particularly bad. When the indicators of both global economy and industry are well, the investor will look into the level of individual companies to find interesting ones.

Macroeconomic analysis

The fundamental analyst following a top-down approach starts with a macroeconomic analysis. He looks into indicators of the global economy such as inflation, interest rates, exchange rates or GDP[1]. This is important because there is a strong link between the growth of the global economy and the growth of companies.

Governments and third parties regularly release new information of the GDP or changing rates. For investors it is important to follow these evolutions. If the indicators show that the economic climate is good for investing, investors can move forward to the industry analysis.

Industry analysis

Each industry is different and has its own customer base, growth, market share distribution among firms, competition between firms and way of doing business [27].

[1] Gross domestic product: value of all final goods and services produced within a nation in a given year [12]

When a company has a limited customer base, for instance because the government is their biggest client, you should be careful when investing. A choice for a competitor by the big client could have a serious impact on the company economic results.

If a company has a large market share, it will do everything to protect its share. The bigger the share of a company, the better it can spread its fixed costs and this gives it an advantage over competitors. Another advantage of having a large market share is that the company can decide over price issues: they can put the standards for customers and suppliers.

The industrial growth and evolution is also an important factor. Is the industry growing or shrinking? Governments can also create regulations that completely change industries. Regulations can cost millions to companies, for instance the testing of new drugs.

Understanding how an industry works gives the investor a deeper understanding of how companies in that business work. Therefore they should closely follow the reports of industrial groups and industry followers. Based on this information, the investor can see which companies are financially healthy and which companies should be avoided when investing. The investor can then move on to the company analysis phase.

Company analysis

After choosing the right industry, the investor can look to the companies into this industry [27]. He has to examine the different business models of the companies so that he can understand what the companies are really doing. This gives him insight in how the companies work and why he should invest in them.

From the outcome of the industry analysis, the investor will also know which companies are the stronger competitors in the market. Stronger competitors are more interesting to invest in.

A last important thing that an investor should know, is how a company is lead. The investor should know if the management is reliable or the opposite, namely unpredictable. The way a company is managed has a large impact on the stock exchange. Investors should also know how the companies relate with their share- and stakeholders.

Based on the company analysis, investors can see which stocks are priced to high or to low and then they can decide over buying or selling shares.

2.4.2 Bottom-up approach

The bottom-up approach starts by analysing individual companies, regardless of their industry. Investors working in this way are just looking for companies with good prospects. They will do the company analysis of section 2.4.1 without doing a deeper research on cross-industry advantages or disadvantages of investing in a given industry.

2.5 Technical analysis

Technical analysts try to predict the movements of the stock by basing themselves on historical data of stocks: prices and volumes. They do not rely on economic insights. There are several ways to do this prediction: based on chart patterns or based on technical indicators. Technical analysts try to foresee the changes in supply and demand, and try to make profit out of this information.

Technical analysis is the opposite of fundamental analysis. Fundamental analysis takes a long-term approach, e.g. some years, while technical analysis is more short-term, some weeks, days or even shorter. Fundamental analysts try to find stocks to invest in, while technical analysts are looking for stocks that they can trade later at a higher price [28].

Technical analysis and the EMH are in contrast with each other. The weak form of EMH assumes that stock prices only reflect historical data. All other information is not included in the stock price. Technical analysts, by contrast, think that they can anticipate on news articles. They believe that if investors wait for news before they invest, they will come too late and will loose the opportunity.

Technical analysts make 3 basic general assumptions, which are explained in section 2.5.1. Some of the strategies they use to make their technical analysis are explained in section 2.5.2.

2.5.1 Assumptions

Market action discounts everything

Technical analysts assume that the stock prices and volumes reflect everything that affected the company. This includes fundamental and economic factors or market psychology. Technical analysts believe that it is not necessary to study those factors separately as they are present in the stock prices.

Prices move in trends

Technical analysts believe that stock movements occur in trends. They assume that movements in certain directions will continue in the future once a trend has been established.

History tends to repeat itself

Technical analysts believe that the history repeats itself. They assume that the market has a repetitive character and that market participants have a consistent reaction on certain events. Technical analysts can then use the chart patterns of the past to try to predict the future.

2.5.2 Strategies

Technical analysts can use different strategies to do their analysis. Two of them, the use of chart patterns and the use of technical indicators, are explained in this section.

Some studies have shown that combinations of multiple strategies perform better than focus on one technique.

Chart patterns

Technical analysts use chart patterns to discover buy or sell signals from the time-price charts of stock prices. They rely on the third assumption of technical analysis, namely that history will repeat itself. They assume that patterns from the past can help identifying trading opportunities when trying to find new trends. However, it is not possible to determine with 100% certainty what will happen.

There are two types of patterns within this domain: reversal and continuation. A reversal pattern signals that a prior trend will reverse upon completion of the pattern. A continuation pattern, on the other hand, signals that a trend will continue once the pattern is complete [28].

In [28] some examples of patterns are presented: head and shoulders, cup and handle, double tops and bottoms, triangles, flag and pennant, wedge, gaps, triple tops and bottoms, rounding bottom. There are known patterns that occurred multiple times in the past and that can help to understand current patterns.

Technical indicators

Some technical analysts use technical indicators to predict the stock movements. Indicators are mathematical formulas with stock information, such as the stock prices and traded volumes, as input. Indicators can be used to confirm trends, to determine the quality of chart patterns or to form buy and sell signals [28].

Technical indicators add an extra layer of information for investors. Based on these technical indicators, an investor can build a trading system and use it as a real trading application. He should follow the outcomes of the system, even if it makes him temporary loose money. The idea is that the system will only be profitable on the long-term and the investor should not have doubts on it.

Examples of technical indicators include money flow, trends, volatility, moving average and the relative strength index.

2.6 Conclusion

Financial markets are markets where people and entities can trade financial assets. There are different kinds of markets, but the one of interest in this master thesis is the stock market. The stock market allows investors to buy and sell equity securities from publicly traded companies.

Predicting the stock market is the big dream of every investor. Nevertheless, researchers do not agree upon the possibility of predicting the stock market. This chapter described the theoretical background of predicting the stock market by explaining the random walk hypothesis and the efficient market hypothesis. This results in two opinions in how to do the prediction, some believe in fundamental analysis, while others believe that technical analysis is the answer.

Fundamental analysts try to use economic parameters to do their analysis. They can do this in two ways: top-down or bottom-up. Top-down analysis starts with inspecting the global economy, then continues by exploring interesting industries and finishes by finding the right companies. Bottom-up analysis starts at the company level and does not give to much attention to global or industrial factors.

Technical analysts try to predict the movements of the stock market by looking to historical data of stocks: prices and volumes. They rely on three assumptions: market action discounts everything, prices move in trends, and history tends to repeat itself. Technical analysts use different strategies to find out how the stock is moving, for instance by using chart patterns and technical indicators.

Chapter 3

Sentiment analysis

Sentiment analysis is the task of detecting, extracting, and/or summarizing opinions, polarity and/or emotions, normally based on the presence or absence of sentiment features [36]. Sentiment analysis tries to find the attitude of the writer or speaker towards the topic he is talking about.

Sentiment analysis could be seen a classification task. The most basic task is determining whether a text expresses an opinion or a fact. Facts are neutral, whereas an opinion can be positive or negative. The task become more complex when finer levels of emotions are used, for instance the six basic emotions as defined in [15]: anger, disgust, fear, happiness, sadness and surprise.

Tasks that perform sentiment analysis have typically three parts. The first subtask is to find proper data. This is described in section 3.1. The next step is to define the sentiment features. Section 3.2 explains this with more detail. The last part is to build an algorithm to do the actual analysis. Section 3.3 goes into deeper on this part. The chapter ends with a conclusion.

3.1 Data preparation

The first thing needed in sentiment analysis is to collect a good data set. Collecting opinion containing documents is a rather easy task. There are lots of web pages containing reviews. Those reviews contain opinions and are often coupled to scores, which give a quantitative indicator of how positive or negative an opinion is.

For some approaches, such as supervised classification, an annotated data set is necessary. The annotation can be done manually or automatically. Manual annotated data is often of higher quality, but also labour intensive. Automatically annotated data can be noisy, however this might not be a big problem in large data sets.

In any case, a gold standard or ground truth is necessary for the evaluation of the system. When comparing the results of the sentiment analysis with the gold standard, conclusions can be made on the research hypothesis.

3.2 Sentiment features

The literature on sentiment analysis provides four feature categories that can be used to help determining the sentiment: syntactic, semantic, link-based and stylistic features [2]. The following sections go more into detail in each of those four possibilities.

3.2.1 Syntactic features

Unigrams

Unigram features, or bag-of-words, finds its origin in natural language processing and information retrieval. This model uses the frequency of words in a training set as features. The frequencies of the words are often normalized by a weighting factor, for instance the document-length or the inverse document frequency. The weighting factor can also be independent from word order or grammatical rules.

Sentiment analysis will now use these word frequencies to build models of probabilities. Certain words have a higher frequency in documents with a positive sentiment, whereas others have a higher occurrence in documents tagged as negative.

Unigrams are useful to use as a basic classifier to make the distinction between positive or negative texts, but it is less useful to make the difference between sentiment and non-sentiment. Therefore, bag-of-words can only be used as a baseline to estimate the difficulty of the target domain as well as providing a basis for the fusion of features [36].

N-grams

N-grams are an extension of the unigram model by taking *n* words together. The big advantage of this is that it preserves context, because words that appear together in the text will also appear together in a n-gram.

Using n-grams can increase the accuracy in classifying positive and negative opinions. Cui et al. show this in their study [13] on sentiment classification for online product reviews.

A special case of n-grams are collocations. Collocations are combinations of words that occur more together than they would by chance. Moreover, the meaning of the collocation is also totally different than the sum of the meanings of the individual words. Examples of collocations are "United States" or "chief executive". It is clear that losing the meaning of this collocations by using unigrams is detrimental for the quality of sentiment analysis.

Part-of-speech tags

Part-of-speech (POS) tags give the linguistic category of words. Examples of POS tags are nouns, verbs or adjectives. Part-of-speech tags can reveal the sentiment evidence of certain groups of words or phrases.

Adjectives or adverbs have a higher occurrence in texts that express an opinion and a lesser occurrence in fact-based texts. POS can clean the uncertainty when using unigrams alone. Some words have a strong sentiment when used as adjective, but are neutral when used as adverb. An example of this is the word "pretty".

First and second person pronouns may also express more often a subjective opinion than objective facts. Though, studies do no mutually agree in this matter. Some, such as [74], use personal pronouns to improve their results in classifying blog posts, whereas studies as [10] are more reserved in their enthusiasm, showing a slightly negative effect. One way to overcome this problem is by using personal pronouns in combination with other features.

Other POS features as verbs and nouns can also be used for sentiment analysis. Verbs can give indications of polarity. In his study [10] Chesley shows that asserting and approving verbs, are strong indicators of positive sentiment. Nouns are more difficult to test, because it is hard to make good lexicons.

Phrase patterns

Phrase patterns make use of POS-tagged n-grams. Fei et al. show in their study [18] that when the subject of review is followed by positive adjectives, the review is almost always positive, whereas the subject of review followed by negative adjectives indicate negative reviews.

Whitelaw et al. use in their study [62] the concept of appraisal groups. This appraisal groups consist of a head term accompanied by some appraisal modifiers. An example is "very famous" where "famous" is the head term and "very" the modifier. The study shows good results for the classification of movie reviews.

3.2.2 Semantic features

Semantic analysis tries to find the correlation between semantic concepts and positive or negative sentiment. The rational behind this is that certain concepts will have a consistent link to a specific polarity. This can help to determine the sentiment of similar semantic concepts and hence improve the accuracy of sentiment analysis [45].

Semantic analysis can be done by manual or (semi-)automatic annotation techniques to add scores to words or phrases. Manual or semi-automatic lexicons are typically composed of automatically generated terms which are later filtered and coded manually with a certain polarity. Different studies, such as [67] or [18] follow this approach.

Riloff et al. use in their study [44] also a semi-automatic generated lexicon to distinguish subjective from objective phrases. In order to do this, they not only consider the meaning of the sentence itself, but also the meaning of the text surrounding the sentence. They give a subjectivity and objectivity score to the sentences before and after the sentence that is studied.

3.2.3 Link-based features

Link-based features use link analysis to determine the sentiments for web documents. Link analysis is a rather new domain of research in this context, so it is not clear how effective it is to help in the sentiment analysis task.

Efron found in [14] that web pages sharing the same opinion are heavily linking to each other. Efron tested his hypothesis on political blogs and music web sites.

Agarwal et al. showed in their study [3] that replies in newsgroups tend to be antagonistic. What people mostly do when replying in newsgroups, is copying the part of another post and then add their own opinion on it. This opinion is often the inverse of the opinion of the first author.

More research is needed to get more insights in how link-based features can be used in sentiment analysis.

3.2.4 Stylistic features

Stylistic features are another domain of research within sentiment analysis. Stylistic analysis makes interpretations of texts based on the linguistic and tonal style. This stylistic analysis can be independent of the topic.

In their study [63] Wiebe et al. tried to find a relation between stylistic features and subjectivity. They found that unique words are a good indicator for subjectivity and opinions. People tend to use more unique words when writing subjective texts.

Gamon shows in another study [21] that deep linguistic analysis features can contribute to the sentiment classification task. The linguistic analysis in his paper is done by NLPWin, a natural language processing system developed by Microsoft Research. NLPWin provides a phrase structure tree and a logical form for each string, from which additional features can be extracted. The domain he uses for his study is customer feedback, a very noisy domain. He also indicates that more research is needed to get more information about the applicability on other domains.

As with link-based features, more research is needed to reveal all possibilities of stylistic features.

3.3 Algorithms

After having determined what features to use, a good algorithm is needed to perform the sentiment analysis itself. There are two possible ways of assigning a sentiment: classifying or giving a score. Classifying is typically done in one of three categories: positive, neutral or negative. If required, finer grained sublevels can be introduced. If the algorithm determines a score, then this score is typically a measurement of how likely is that a text belongs to a certain category.

This section makes the distinction between three kinds of methods to solve the problem: by using an ad hoc rule-based approach, by using a supervised learning approach or by using a semi-supervised learning approach [36].

3.3.1 Ad hoc rule-based approach

Rule-based systems use rules to make deductions or choices. In this case the rules are a list of terms or patterns that provide evidence for the presence of sentiments or facts. The most basic systems use matching rules that couple one or more sentiment features to sentiment classes.

The advantage of rule-based approaches is that it is intuitive and easy to understand and implement. Despite this advantages, there are also some difficulties. The sentiment features must be of good quality and they must allow to correctly determine the different sentiment categories. Accuracy of classification will be low if the features are of poor quality or if they are not diverse enough. Rule-based approaches have also difficulties with exceptions, because they can never capture all exceptions [36].

One last issue with rule-based systems is that they require quite a lot of manpower to compose. People need to have a good knowledge of sentiment features before building such a system. Moreover, manpower must also be invested on maintaining and updating the system.

Rule-based systems are widely spread and applied in commercial systems. Brandwatch[1] is an example of such an application. It tries to figure out the reputation of brands. It does so by using a library with hundreds of rules. These rules are made by analysts who marked up thousands of mentions in different languages [6].

3.3.2 Supervised learning approach

Supervised learning is a machine learning task that tries to infer a function from labelled training data. Supervised learning has proved itself in the domain of topic classification and is also useful in the domain of sentiment analysis.

Two of the most popular supervised learning approaches, naive Bayes and support vector machines, are explained with more detail in this section.

The biggest difficulty in using supervised learning methods is the need of qualitative training data. Supervised learning is also sensitive to biased data. The performance of the system might suffer from this.

Naive Bayes

Naive Bayes is a probabilistic classifier. It assumes strong independence between the different features and uses Bayes' theorem. It is very popular because of its simplicity. It works well in classification tasks, so it is also used in sentiment analysis.

Prasad shows in his work [43] that naive Bayes performs well in the classification of polarity of microblogging posts on Twitter. Gamallo describes in his work [20] the use of naive Bayes to do the same classification on Spanish tweets.

Pang et al. compare in their study [40] different machine learning techniques and found out that naive Bayes is not performing bad, but it scores generally worse than some other techniques. This observation has been confirmed by several other

[1] http://www.brandwatch.com/

researches: naive Bayes performs good but other techniques, such as support vector machines, are better.

Support vector machines

Support vector machines (SVMs) are used for classification and regression tasks. SVMs use a set of training examples as input set. Each of those training examples is assigned into one out of two categories. A SVM builds a model that represents these examples as points in space and tries to separate the categories by a clear gap that is as wide as possible. New examples are mapped to the same space and a category is assigned based on which side of the gap they fall on.

Different studies use SVMs in sentiment analysis. Yang et al. use for instance SVMs to do sentiment classification on text largely crawled from internet sources [69]. Other studies try to improve the basic SVM by choosing special features. Mullen and Collier try to improve SVMs by adding topic information [37]. SVMs using features containing that topic information outperform models that don't. Whitelaw et al. use appraisal groups as input features for their SVM [62]. This gives better results than SVMs that don't use appraisal groups.

3.3.3 Semi-supervised learning approach

Semi-supervised learning use not only a labelled input set, but also some unlabelled data. The reason to do this is that in many applications only a limited amount of labelled data is available. Getting labelled data is expensive because it requires quite some manpower, while unlabelled data can be collected cheaply and automatically.

Many researches have found that when there is only limited labelled data available in the data domain, semi-supervised learning can achieve improvement over supervised learning.

Semi-supervised learning approaches rely on three assumptions: the smoothness assumption, the cluster assumption and the manifold assumption. Those assumptions are deeper explained in the work of Chapelle et al. [8].

This section explains some important semi-supervised methods that can be used in the context of sentiment analysis.

Self-training

Self-training strategies start with an initial sentiment lexicon as seed and then use iterative training to enlarge that lexicon. Zagibalov and Carroll employ this technique in [71] using a seed word (for instance "good") to iteratively create a training set. They get good results in the classification of product reviews in Chinese.

Riloff et al. use the same technique in their study [44] to make a semi-automatic generated lexicon to distinguish subjective from objective phrases. They start with a collection of unannotated texts and a few seed words, and then learn a lexicon of subjective nouns.

He and Zhou begin in their study [24] with a set of unannotated texts and a sentiment lexicon. This sentiment lexicon contains a list of words and their polarity. They use self-training to iteratively learn a set of self-learned features.

Expectation-Maximization

Expectation-maximization (EM) algorithms try to estimate the maximum-likelihood for models with incomplete data. The EM algorithm starts with the expectation (E) step, where it creates an expectation function for the parameters it has to find. This is followed by the maximization (M) step, where the algorithm tries to compute the parameters by maximizing the expected likelihood from the E step. After the M step, the algorithm repeats itself iteratively until a stop criterion is reached.

EM-based approaches are also applied on text classification problems, where unlabelled data is treated as incomplete data. Most applications use an EM algorithm with a naive Bayes classifier. The study of Buche et al. [7] uses this approach in the analysis of online reviews of customers on products.

Zhang et al. point in their work [73] that there might appear a problem in the standard EM procedure: it may optimise parameters of a generative model whose objective function does not conform to its intented purpose of classification. This may lead to poor classification performance. They solve this by extending the EM algorithm with the lexical knowledge of word labels. This improves the E step in the algorithm and extends it by combining the document class labels with labels created from word labels. They claim to have better results than when they build models only based on the labelled data.

Semi-supervised support vector machines

Semi-supervised support vector machines (S3VMs) are an extension of standard SVMs. They try to find the maximized gap between two categories by using both labelled and unlabelled data. As all other semi-supervised methods, it has the big advantage of requiring less labelled data and so less human effort.

Setiady et al. uses S3VMs to do sentiment analysis of online reviews [53]. They found out that S3VMs perform slightly worse than ordinary SVMs, but they require 40% less labelled input. So they believe that S3VMs are more interesting than normal SVMs. By using S3VMs, you can reduce time and costs needed to label a data set manually. They report also one disadvantage when using the semi-supervised variant: sometimes the algorithm gets stuck in a local optimum, which slows down the processing time.

Yates et al. combine in their study [70] a lexicon-based approach and a SVM and use this as the baseline of their work. Their results show that semi-supervised lexicon SVMs perform better than lexicon SVMs in classifying online reviews. They introduce another algorithm, based on semi-supervised probabilistic sentiment analysis. This approach performs the best.

3.4 Conclusion

Sentiment analysis tries to find out the attitude of the writer or speaker towards the topic he is talking about. Sentiment analysis tasks consist typically of three parts: collecting a good data set, choosing the sentiment features and running an algorithm to do the actual sentiment analysis task.

Good data is a necessary starting point for sentiment analysis. Collecting documents containing opinions is not a difficult task. Depending on the algorithm that is used, an annotated data set can be necessary. This labelling is often a time-intensive and money-consuming task, because it needs human supervision.

Sentiment features help to determine the sentiment. There are several categories of features: syntactic, semantic, link-based and stylistic features. Each of those feature categories has examples in literature that show that they perform well, however sometimes a combined approach would perform better.

Algorithms performing sentiment analysis can be divided in three categories: ad hoc rule-based approaches, supervised learning approaches and semi-supervised learning approaches. The advantage of the first is that it is easy to understand and implement. The disadvantage is that it largely depends on the good quality of features. Supervised learning approaches have proved themselves in classification, but need a large annotated data set. This is where semi-supervised approaches can help: it needs only a small data set and will infer features for unannotated data.

Chapter 4

Existing systems

The relation between stock quotes and textual data has fascinated researchers since the end of the 20th century. In 1997 the first research projects on this subject started at the Hong Kong University of Science and Technology. Soon different other institutions followed their example. This chapter describes some interesting projects in the area.

Section 4.1 focusses first on projects that used textual data coming from (financial) news articles. It gives a chronological overview of interesting projects and milestones. The section starts in Hong Kong and ends with the latest state-of-the-art research projects.

During the years, different other sources of textual data appeared. An example of this are financial reports. Section 4.2 describes a system that uses these reports as input.

Another data source is the topic of research in this master thesis: Twitter. Its booming popularity attracted researchers to leave the financial news articles and to investigate if Twitter influences the stock market. Section 4.3 presents in a chronological order some research projects that use Twitter as data source.

The chapter ends with a short summary of all existing systems.

4.1 Approaches based on news articles

4.1.1 Early work in Hong Kong (1997-1998)

The earliest work in this domain was done at the Hong Kong University of Science and Technology. First, two master students wrote their master thesis on this topic, followed by other work at the same university.

The first student, Desamanya Peramunetilleke, wrote in 1997 his master thesis [41] on a system for exchange rate forecasting using news headlines. He used tuples of keywords coming from news titles. His system uses probabilistic rules to forecast three categories of exchange rate movements: up, down or steady.

The second student, Steven Leung, wrote in the same year his master thesis [31] on automatic stock market predictions from World Wide Web data. In his thesis he investigated the possibility to predict the closing values of the Hang Seng Index (HSI), the most used stock market index of Hong Kong. He used keyword pairs, triples and quadruples from electronic newspapers. Again the systems uses probabilistic rules to predict the movements of the stock market using three categories: up, down or steady.

One year later, in 1998, Cho et al. continued the research and wrote a work [11] on text processing for classification. They used real-time news sources in their work, containing global and regional political and economic news, citations from bankers and politicians, and recommendations from financial analysts. They tried to forecast the movements of the HSI, just as Leung, by using manually selected keywords. They end their work with suggesting that a clustering approach could be useful to cluster articles with the same meaning and exploit their common features to predict the stock market.

4.1.2 Lavrenko et al. (2000)

Description

Lavrenko et al. describe in their paper [29] a prediction system called Æneast. Contrary to the work described in the previous section, this system tries to predict individual stocks and not an index of combined stocks.

Their system first identifies trends in the stock data time series by using piecewise linear fitting. Those trends are discretized in five categories: big rise, small rise, no effect, small drop and big drop. Then it selects relevant news articles for the involved stock. The news articles are coming from Yahoo Finance. Then the articles are linked to the found trends and the system trains for each of the trends a language model. The used language model is a naive Bayes classifier.

Finally, fresh published news articles are compared to the language models of each of the trends. This results in a likelihood on each of the trends and hence a global probability that a certain trend will occur.

The time lag that was used between the articles and the stock movement is ranging from 1 to 5 hours. The stock prices had a granularity of 10 minutes.

Results

Their system demonstrated that it could successfully associate news stories to stock trends by using language models.

They also built a stock market simulator that used Æneast. That simulator made 280000 dollar profit, but without taking transaction costs into account. This is a commonly followed approach in papers, however the transaction costs are often very high. The system did more than 12000 transactions in 40 days, which is a very high number and therefore very expensive.

4.1.3 Gidófalvi (2001)

Description

Gidófalvi presented in 2001 a new model [22]. This model uses the same granularity of stock data as Lavrenko, namely 10 minutes. Another similarity to the work of Lavrenko is that they use a naive Bayes classifier to do the classification. They use three categories in their classification: upward movement, downward movement and expected movement.

The most important novelty is the search for a window of influence. This window of influence is the time period for which news articles have an impact on stock prices. In their work different windows are tested and compared.

Results

Gidófalvi found that the best relative prediction accuracy occurs for windows of influence that open 20 minutes prior and close 20 minutes after news articles become publicly available. This is in contradiction with the efficient market hypothesis, which says that all information is contained in the stock prices.

However, even if their results are significant, the predictive power of their system is limited. Gidófalvi pointed out two reasons for this. The movement measure could be improved by using more realistic index price information instead of an approximated version. The second reason could be found by the duplication of news by the different news agencies: only the first appearance of news is probably important for the stock influence.

4.1.4 Fung et al. (2002)

Description

Fung et al. present in their paper [19] a model that is based on the efficient market hypothesis, as introduced in section 2.3.3. They try to predict the stock market behaviour based on non-quantifiable information. They obtain this information from news articles.

Similar to Gidófalvi, they try to cluster stock market trends into one of three classes: rise, drop or steady. Then they introduce two new algorithms in this domain: guided clustering and a new weighting scheme. They use incremental K-means for the alignment of news articles with the discovered trends. A bag-of-words representation is used for the news articles, with td-idf weighted features.

Finally, two SVMs are trained based on the associations between the different trends and features. One SVM is used to check if news articles trigger a rise of the stock market or not, the other one does the same for drops in the stock market.

Results

They tested their system for 614 stocks of the Hong Kong exchange market during seven months. During this period they collected 350000 news articles.

The receiver operating characteristic (ROC) curve of their system showed that their system performs better than random and non-guided approaches. Their research made also clear that the amount of news articles is important. The more frequent that news articles appear, the better an alignment of news articles with stock trends can be made.

4.1.5 Schumaker et al. (2006-2010)

Robert Schumaker wrote different papers on the Arizona Financial Text System (AZFinText), a research project that uses financial news articles to predict future stock prices. The system was introduced for the first time in 2006 by Schumaker and Chen and was used during the following years to answer different research questions that examined stock market prediction. This section describes his work chronologically.

Schumaker & Chen (2006)

The first paper [48] on AZFinText appeared in 2006. In this paper Schumaker and Chen use a window of influence of 20 minutes, as earlier introduced by Gidófalvi. Their system is used to determine which textual analysis technique is the most valuable for stock price prediction. The three used techniques to represent the news articles are the classical bag-of-words representation, noun phrases and named entities. Then they train three different SVM models. The first uses only the different article representations for its prediction. The second uses the representations and the stock price on the moment that the articles appeared. The third uses the representations and a regressed estimate of the stock price 20 minutes after the moment that the articles appeared.

They find that the model using both the representation and the stock prices on the moment of appearance performed the best. It is able to predict the stock prices better than the other models. The best textual representation technique is the named entities, followed by the noun phrases. The bag-of-words representation is performing poorly. The three metrics used to determine the best technique were closeness, directional accuracy and simulated trading.

Schumaker & Chen (2008)

A second paper [49] followed two years later. In this paper they try to build a hybrid quantitative system that incorporates both traditional quantitative trading strategies and financial news prediction. In quantitative investing, stocks are chosen based on some considerations: social responsibility of companies, diversification issues (a mix of companies of different sectors), investment risk, effect of taxes, growth, etc.

In their work they test four models of varying levels of quantitative and financial news prediction as well as different portfolio formation periods. The portfolio formation time is the period where returns are analysed and stocks are selected.

Results show that a hybrid approach performs the best. It is necessary that all news articles are used as input for the model, not only the news articles of the companies from the quantitative portfolio. Their results show also that returns lower with additional portfolio formation time.

Schumaker & Chen (2009)

The third paper [52] of Schumaker and Chen builds further on their first paper. It describes again experiments with the three textual analysis techniques as presented earlier. They introduce in this paper a forth technique: proper nouns. Proper nouns are a subset of noun phrases and an intermediary between noun phrases and named entities.

Their experiments show that proper nouns perform better than their parent, noun phrases. Named entities come forward as the best technique in the first paper, but now proper nouns perform better in two of the three used metrics. The authors suggest that the success of proper nouns can be declared because they are freer of the term noise used by noun phrases and free of the constraining categories used by named entities.

Schumaker & Chen (2009)

In 2009 another paper [50] of Schumaker and Chen was published. This paper tries to find out if the usage of the Global Industry Classification Standard (GICS) has a positive influence on the prediction of the stock market. The paper also compares the forecast of AZFinText to the predictions of human trading professionals and quantitative funds.

GICS is composed of different classification levels: universal, sector, group, industry, sub-industry, stock specific. All those levels are compared with the three metrics that they use in every paper. Sector-based training performs better than other classification levels.

The results of this paper show the sector-based approach performs the better than 2 trading professionals, the first place is obtained by DayTraders.com. In simulated trading, AZFinText outperforms six of the top ten quantitative funds.

Schumaker, Zhang & Huang (2009)

Schumaker et al. wrote one month later a paper [51] in which they performed sentiment analysis on news articles. Two research questions appeared: does objectivity/subjectivity impact news article prediction and does positive/negative subjectivity impact news article prediction.

The first finding in this paper is that subjective articles perform better in terms of accuracy and trading results than the model used in previous papers, but not closeness. Using a tone model, which is able to determine subjectivity/objectivity does not improve the results. Subjective articles influence market trading immediately following the article release time, there is no time lag between new articles and the reaction of the stock market.

Negative subjective articles seem to perform the best when comparing to the baseline and using accuracy and trading results as metric. The closeness of the baseline model remains the best. This can be declared by the strong reaction of investors when reading negative articles.

Schumaker (2010)

In his next paper [46] Schumaker tries to figure out the role of verbs in financial news articles. A feature vector of each news article is made based on the presence of verbs, with a feature 1 if a verb is present and 0 is a verb is absent. This is then used as an input vector for a SVM, combined with the stock price on the time the article was published.

Their experiments result in a formula which predicts the stock price, based on verbs used in the articles. However, even if they always use the word "verb" in the paper, some of the words coming out of their experiments are not real verbs. The five words with the highest negative impact are "hereto", "comparable", "charge", "summit" and "green". The word "comparable" has for instance a negative impact of 0.002 dollar. The words with the highest positive impact are "planted", "announcing", "front", "smaller" and "crude".

Schumaker (2010)

Schumaker continued the work of his previous paper with a new paper [47] where he tries to find a stock price predicting formula for the different textual representations that were discussed in earlier works: bag-of-words, noun phrases, named entities, proper nouns and verbs. The purpose of this research was to find terms that appear across similar textual representations.

Positive, negative and neutral support vectors are created that have respectively a positive, negative or neutral influence on the stock price. For each of the textual representations a formula is created to predict the stock price, just as in the previous paper [46]. Some terms of the more restrictive textual representations also appear in the bag-of-words representation if enough weight is assigned to them. The words found are mostly specific for the context of the data that was used in this paper, and not useful in other contexts.

4.2 Approaches based on financial reports

4.2.1 Lee et al. (2010)

Description

Lee et al. use in their paper [30] a totally different data source than the papers of the previous section: they use financial reports to predict the stock market. They also introduce a new approach to perform the prediction when comparing to systems from before 2010: a clustering approach.

First they convert each financial report into feature vectors. Those feature vectors contain both qualitative and quantitative features. The qualitative features come from the textual contents of the financial reports, whereas the qualitative are coming from the numbers in the reports.

Then they use hierarchical agglomerative clustering to divide the converted features into clusters. The clusters are classified with a tag to indicate if the stock market rised, dropped or stayed equal the day after the publication date of the report. For each cluster they then recursively apply the k-means algorithm to divide it further into sub-clusters, each with a centroid.

New financial reports are finally compared to the different centroids to determine the prediction of the stock market based on the tag (rise, drop or steady) that the centroid has.

Results

Their results show that their hybrid approach of hierarchical agglomerative clustering and K-means works, and that their approach seems to perform better than SVMs. Lee et al. show in this way that using a clustering algorithm can be a good alternative for the use of SVMs.

They also found that a combination of qualitative and quantitative features in financial reports performs better than only considering one of the two.

4.3 Approaches based on Twitter

4.3.1 Pak & Paroubek (2010)

Description

Pak and Paroubek built in their research [39] a sentiment classifier that is able to determine positive, negative and neutral sentiments. They use microblogging posts from Twitter as their data source.

In their research they use a multinomial Bayes classifier, because it gives the best results for them. It scores for instance better than SVMs. The classifier uses n-grams and POS-tags as features. The POS-tags are tagged by TreeTagger[1], the n-grams vary in size during the experiments.

Results

Pak and Paroubek try their classifier for different sizes of n-grams. Their experiments show that bigrams are performing the best, because they provide a good balance between both a good coverage of the data (as unigrams do) and a good capture of patterns of sentiment expressions (as higher n-grams do). The accuracy can be even improved when taking the negation of words as one word. To give an example: the word "do" followed by "not" would count as one word do+not of the bigram.

They also show that some POS-tags are stronger indicators of emotional text than others. Objective texts contain more common and proper names. The verbs in

[1] http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/

objective texts are used more often in the third person or in past participles. At the other side of the spectrum, utterances and personal pronouns are strong indicators of subjective texts. Authors use verbs in the first person (to talk about themselves) or in the second person (to address the readers) in subjective texts. Emotions are expressed in the simple past tense or by using modal verbs. Superlative adjectives are more used for expressing emotions and opinions, while comparative adjectives are more used in facts.

Increasing the data set size increases the F-measure and hence the performance of the system. However, at a certain point the increase of training data is not enough to achieve better performance.

4.3.2 Bollen et al. (2011)

Description

Bollen et al. try to find out in their work [5] if a society can experience mood states that affect their collective decision making. Or more specific, if the public mood can be used as an economic indicator.

They analyse the mood in large-scale Twitter feeds and try to correlate it to the Dow Jones Industrial Average (DJIA). For obtaining the mood they use two tools: OpinionFinder and Google-Profile of Mood States (GPOMS). OpinionFinder measures positive versus negative mood, while GPOMS measures mood in six dimensions: *calm*, *alert*, *sure*, *vital*, *kind* and *happy*.

They use granger causality analysis and a self-organizing fuzzy neural network to investigate the hypothesis that public mood states are predictive of changes in DJIA closing values.

Results

Bollen et al. create first series of mood states for tweets posted between October 2008 and December 2008. Then they try to find correlations between the time series of OpinionFinder and the six time series coming from GPOMS. They find that the OpinionFinder series was significantly correlated with the series of *sure*, *happy* and *vital*.

Then they try to find correlations between the mood series and the DJIA series. Therefore they use a longer period: from February 2008 until November 2008. They find that the *calm* mood of GPOMS is significantly correlated with stock changes with lags from 2 to 6 days. All other series were not correlated to the DJIA.

They conclude their work with trying to predict the stock changes based on the different moods that they obtain. Their research concludes that a prediction of the direction is better when it is based on the *calm* mood series. The prediction gets more accurate (with a lesser mean absolute percentage error) when a combination of *calm* and *happy* is made, but the result is worse than *calm* alone. They conclude that there must be a nonlinear relation among the different mood dimensions.

4.3.3 Chen & Lazer (2011)

Description

Chen and Lazer use in their paper [9] a pre-generated word list of 5000 common words. Each of those words has a probability of being associated to "happy" and "sad". According to the words that are used in the tweet, a total probability of the tweet for being associated with "happy" and "sad" can be made. They divide all obtained sentiments per day by the number of tweets to get an average sentiment.

They also have a variant on this first algorithm where they use SentiWordNet[2]. This uses a more comprehensive and accurate dictionary (consisting of 400,000 words) for positive and negative sentiments. It also considers relations between words and multi-word expressions.

They design several features that correspond to the sentiment values of tweets. They also add some features to model the changes in price of the market each day and the total change in the past n days. All those features are modelled with various time-delays.

According to these features, a prediction of the stock market price is made. Two investment strategies are used, each based on that prediction. The classification strategy just looks to the sign of the prediction: if it is positive, shares are bought, if it is negative, share are sold. The regression strategy tries to exploit the knowledge of how much the stock will change. It allows to invest on a more differentiated way.

Results

The use of SentiWordNet gives better results than the use of the pre-generated list of 5000 words because it is more differentiated. They discover that the best results are obtained when the Twitter data precedes the market data by about 3 days.

Both investment strategies result in more profit than the default buy and hold strategy. The classification model is more consistent, while the regression strategy is more sensitive to noise. The more data was used to train the model, the better the regression model works. Less data results in more imprecision and hence less investment decisions. The classification strategy is more robust to noise.

4.3.4 Zhang et al. (2011)

Description

Zhang et al. try to find the collective people's mood in their research [72]. They do so by counting all tweets containing emotional words. These words can be divided in two groups: positive, as hope and happy, and negative, as fear and worry. They get a constant feed of streams from Twitter, so they can express for each of those emotional words the occurrence as a percentage. They also express the number of followers and the number of retweets of a tweet with a certain mood as a percentage.

[2] http://sentiwordnet.isti.cnr.it/

Then they try to measure the correlation between those percentages and different indices: Dow Jones Industrial Index, NASDAQ Composite, Chicago Board Options Exchange Volatility Index VIX and Standard & Poor's 500. For more information about indices, see section 2.2.3. VIX is negatively correlated to the other ones, this is normal because the spread of stock options is used to calculate VIX.

The researchers also investigate the time lag between the tweets and the prediction of the stock market.

Results

Zhang et al. collected Twitter feeds for 6 months at an average of 29758 tweets per day. The first thing that they find is that the number of positive tweets is much higher than the number of negative ones, more than double on average. This might indicate that people prefer optimistic words when writing tweets.

They find a positive correlation between the number of tweets and VIX and a negative correlation between the number of tweets and the other indices. This implies that people start using more emotional words in times of economic uncertainty. The polarity of the words is not important in this context.

They find that the number of followers for tweets with a specific mood is not a good predictor of stock market indices. They also find that the number of retweets is a good indicator for the total activity of all Twitter users, but that it is not a good predictor of stock market indices. The number of tweets is a better indicator than the number of retweets.

They find that different time lags gave the same results: the use of emotional words is negatively correlated with DJIA, NASDAQ and S&P 500 and positively with VIX. They test time lags of one, two and three days. They get the best results for the words "hope", "fear" and "worry".

4.3.5 Evangelopoulos et al. (2012)

Description

Evangelopoulos et al. examine in their work [16] if an aggregate of Twitter messages can be used as a predictor of future stock prices. They use latent semantic analysis (LSA) to extract semantic and conceptual content in the form of key themes. They built a regression model that uses tweet volume and tweet topic strength to predict the stock prices.

Results

They first collected a data set of tweets of 18 Fortune 500 companies. Then they take a sample of the tweets of each company, with each sample representative for the total number of tweets for that company. On this smaller data set they perform LSA. They make a 20-factor solution: a list of 20 high-loading terms.

These factors can be divided in several themes. A first group of factors formed around types of products and services, such as "laptop" or "phone", but also more specific as "Microsoft Office". The second group is related to shopping. Examples of this group are "gift card" or "sales". Other groups are linked to actions ("buy" or "get") or emotions ("LOL"[3]). Other groups are more stand-alone factors: "job" and "new". Two others, "like" and "all", did not represent any theme.

Then they try to link both the tweet volume and tweet topics to stock changes. They find that the total tweet volume is negatively related to the stock performance during the same trading day. The amount of tweets of some tweet topics have a positive effect on the stock performance of a company, while others have a negative influence. For all topics there is a time lag between the volume of tweets in a topic and the reaction of the stock market. This varies from 1 to 6 days. Tweet topics can explain over 8% of variance in stock performance above that explained by market.

4.3.6 Smailović et al. (2013)

Description

Smailović et al. describe in their paper [54] their research on whether Twitter feeds, expressing public opinion concerning companies and their products, are a suitable data source for predicting the stock movements.

They use a linear support vector machine (SVM) algorithm for the classification of sentiments. They motivate their choice by saying that SVMs are robust to overfitting, they can handle last feature spaces and it is memory efficient.

Then they try to investigate whether the sentiment analysis has a predictive power. They use two categories of experiments: one with just positive versus negative classification and one with an extra neutral class for instances that are close to the SVM's hyperplane.

Results

The classification by the SVM gives good results: it has an accuracy of 81%. This accuracy was obtained by using bigrams, by only using words that appear at least twice in the corpus, replacing URLs by a general token and by removing repeated letters in words.

The first series of experiments, with a positive-negative classifying SVM, were done on a 9 months' time period. Their experiments show that the positive sentiment probability can predict a similar rise or fall of the closing price of companies with a lag of a few days. The correlation is more clear for companies that have many variations in the closing price values or a significant fall in the closing price than for companies that do not have it.

For the second series of experiments a neutral class of tweets was introduced. During the experiments the size of the neutral zone around the hyperplane was variable. Experiments show that adding a neutral zone increases the predictive power of tweets. In general the best improvement is obtained with a neutral zone of 0.2 times the distance to the first positive plus 0.2 times the distance to the first negative

[3] Laughing Out Loud

example. The neutral zone has the best effect when the closing prices show a lot of variations, while it get the worst results when there is a constant fall of closing price.

4.4 Summary

Many researchers searched for relations between textual data and stock quotes. In the beginning this textual data was coming from news articles, whether or not financial. Later other data sources came to the foreground. Two examples of those sources are financial reports and Twitter. In the beginning the prediction systems were primitive, but during the years they got more refined by better insights in the prediction of the stock market.

4.4.1 Approaches based on news articles

The first research projects started in Hong Kong in 1997. The systems used manually selected keywords and probabilistic rules to predict the movements of a stock market index in three categories: up, down or steady. The use of those systems was very limited: human intervention was necessary and individual stocks could not be predicted.

Soon the first improvements on this first systems appeared. Lavrenko built a system Æneast which was able to combine stock data time series and news articles from Yahoo Finance. Æneast used a naive Bayes classifier and was able to predict individual stocks and not an index of combined stocks. The time lag between the articles and the stock movement ranged from 1 to 5 hours.

Gidófalvi made the next improvement. He found the existance of a window of influence: a time period for which news articles have an impact on stock prices. He found that a window of influence opening 20 minutes before the article appeared and closing 20 minutes after, performed the best.

Fung introduced in 2002 SVMs in this research domain. He first aligned news articles and discovered trends in the stock market data. Then he used weighted feature vectors, based on a bag-of-words represenation. His results show that a SVM is a useful tool to predict rises or drops of the stock market.

Schumaker built a system called AZFinText. During the years he wrote different papers, each answering different research questions in this domain. He compared different textual analysis techniques and found that using proper nouns outperformed using bag-of-words, named entities and noun phrases. He found that hybrid approaches of combining news articles and quantitative portfolios perform better than just approaches based on only one of those two elements. He found that a sector-based approach scores better in predicting than approaches using other levels of industry classification. He found that negative subjective articles tend to perform better when predicting the stock market.

4.4.2 Approaches based on financial reports

Lee et al. describe in their paper a system that uses financial reports as input. They uses a hybrid approach of hierarchical agglomerative clustering and K-means to do a prediction of the stock market. Their approach performed better than a SVM and is in this way a good alternative for SVMs.

4.4.3 Approaches based on Twitter

Researchers are attracted to combining stock quotes to Twitter since the popularity of this last one boomed. They hope to find signals in the public mood that help them in predicting the stock market.

Pak and Paroubek built a first system. They adopt ideas from papers that describe predicting systems based on news articles. In their paper they describe a system using a multinomial Bayes classier. The classifier performed the best using bigrams as features. They also found that certain POS-tags are stronger indicators of emotional text than others.

Bollen tried to find out if a society can experience mood states that affect their collective decision making. He uses OpinionFinder and Google-Profile of Mood States (GPOMS) to determine the mood in tweets. He found that the calm mood was significantly correlated with stock changes of the Dow Jones Industrial Average (DJIA).

Chen and Lazer used a pre-generated word list with probabilities to be associated with happy and sad to determine the sentiment of tweets. Later they also did the same using SentiWordNet. Based on those sentiment values of tweets they tried to build a model that could predict the stock market prices. They found that the best results were obtained when the Twitter data precedes the stock market data by about 3 days.

Zhang tried to determine again the collective people's mood. He did that by counting emotional words. Emotional words can be divided in two groups: positive, as hope and happy, and negative, as fear and worry. They found that the number of tweets containing such words is more important than the number of followers for a tweet or the number of retweets of a tweet. They found that different time lags give the same results: the usage of emotional words is negatively correlated with DJIA, NASDAQ and S&P 500 and positively with VIX. The best results were obtained for the words "hope", "fear" and "worry".

Evangelopoulos et al. tried to predict future stock prices by using latent semantic analysis (LSA). They performed LSA and obtained a list of 20 high-loading terms. Then they tried to link the tweet volume of the different topics to the stock price evolution. They found that tweet topics can explain over 8% of variance in stock performance above that explained by market.

Smailović describes in his paper how SVMs can be used for the classification of public opinions concerning companies and their products. He first describes an SVM able to classify positive versus negative and later a second SVM with an extra neutral zone. Results show that adding a neutral zone increases the predictive power of tweets.

Chapter 5

Data collection

As mentioned in chapter 3, all sentiment analysis tasks start with a good data set. This master thesis focuses on sentiment analysis in microblogging posts on Twitter and tries to correlate changes in the sentiment of these posts with changes in the stock market. This means that two sorts of data are necessary: microblogging posts and stock information.

Structured query language (SQL) is a special programming language for managing data. It is used in the context of this master thesis to manage a MySQL database. The features of SQL are explained in section 5.1.1. Section 5.1.2 explains then how the data is structured in the MySQL database.

Section 5.2 explains which existing data sources of microblogging posts where used in this work and how the tweets and users are crawled from Twitter. Section 5.3 explains the used data sets of stock quotes. Stock quotes contain information on the share price and traded volumes on a certain moment in time.

The chapter ends with a conclusion.

5.1 Data storage

5.1.1 Structured query language

Structured query language (SQL) is a special programming language that is designed for managing data in relational database management systems (RDBMS). SQL is a standardised language and is based on relational algebra and tuple relational calculus.

SQL can be seen as the combination of three sublanguages: data manipulation language (DML), data control language (DCL) and data definition language (DDL). The DLM part of SQL provides ways to manipulate data: insert, update, merge and delete. It has also a select method to query data. DCL is used to implement access control on your database. Finally, DDL is used to define the structure of the database, such as fields, tables and indices.

SQL knows different data types: character strings, numbers, bit strings, dates and times. Each of these categories consists of several subcategories, each in turn with their own characteristics.

MySQL is used in this master thesis as RDBMS. It is the second most used open-source RDBMS (after SQLite, which is used on mobile devices). MySQL provides a stable, user-friendly and secure way to store data.

5.1.2 Data structure

The collected data is split over different linked tables. This section explains the different tables, their fields and how they are linked.

Tweets

The *tweets* table contains all collected tweets. It stores every field that the Twitter API returns [59], in particular:

- **ID:** the ID of the tweet.
- **Time:** the time of the tweet.
- **User:** the ID of the user that published the tweet. This field can be used in the *user* table to get the extended information of the user that posted the tweet.
- **Text:** the text of the tweet.
- **Company:** the ticker of the involved company. The full name of the company can be obtained from the *companies* table.
- **Favourite count:** the number of times the tweet has been favourited.
- **Retweet count:** the number of times the tweet has been retweeted.
- **Reply on user:** the ID of the user to whom this tweet is addressed; only present if the tweet is a reply.
- **Reply on tweet:** the ID of the tweet to which this tweet forms an answer; only present if the tweet is a reply.
- **Truncated:** indicates whether the tweet is truncated or not.
- **Possibly sensitive:** indicates whether the tweet has a possible sensitive content or media link in it. It does not screen the text of the tweet, only the content of URLs inside the tweet.
- **Language code:** indicates the machine-detected language of the tweet.

Users

The *users* table contains all collected user information. Only user information from users that posted one or more collected tweets was stored. The table contains the following fields [60]:

- **ID:** the ID of the user.
- **Screenname:** the screen name or alias to which the user identifies himself with.
- **Name:** the name of the user.
- **Description:** the self-defined description of the user.
- **Created:** the date the user profile was created.
- **Location:** the self-defined location of the user.
- **Language:** the self-defined language of the user.
- **Profile image:** the user's avatar image.
- **URL:** a self-defined URL associated with the user.
- **Followers count:** the number of followers the user has.
- **Friends count:** the number of users the user is following.
- **Favourites count:** the number of tweets the user has favourited.
- **Listed count:** the number of public lists the user is member of.
- **Statuses count:** the number of tweets the user made, including retweets.
- **Time zone:** the self-defined time zone of the user.
- **Contributors enabled:** indicates whether other Twitter users are able to co-author tweets on the account of the current user.
- **Verified:** indicates whether or not the user has a verified account.
- **Protected:** indicates if the user has chosen to protect his tweets.
- **Geo enabled:** indicates if the user has enabled the possibility of geotagging his tweets.

Companies

The *companies* table contains background knowledge on the companies. The data inside the table was obtained from a list with the 2000 most valued companies [38]. It has the following fields:

- **Name:** the full name of the company.
- **Ticker:** the ticker symbol of the company, as explained in section 2.2.2.

Sanders twitter sentiment

The *Sanders twitter sentiment* table was named after Sanders Analytics. It contains information from the data collection that will be explained later in section 5.2.1. It has the following fields:

- **Tweet:** the ID of the tweet.
- **Sentiment:** the sentiment that was manually tagged to the tweet. This can be one of the following four sentiments: positive, negative, neutral or irrelevant.

Stock tweets

The *stock tweets* table represents the information of the second data collection of tweets that will be explained in section 5.2.1. It consists of 2.7 million tweets containing a string that starts with a dollar sign ($). It has the following fields:

- **Time:** the time of the tweet.
- **Tweet:** the ID of the tweet.
- **Tag:** the string containing the dollar sign ($).
- **Keywords:** zero or more keywords describing the tweet.

Stock quotes

The stock quotes of a company contain information on the share price and traded volumes on a certain moment in time. They are grouped in different tables, one per company. This is done to keep the sizes of those tables reasonable. All tables contain the following fields:

- **Date:** the date of the stock quote.
- **Time:** the time of the stock quote.
- **Open:** the opening price of the stock quote during the interval.
- **High:** the highest price of the stock quote during the interval.
- **Low:** the lowest price of the stock quote during the interval.

- **Close:** the closing price of the stock quote during the interval.
- **Volume:** the volume of shares traded during the interval.

5.2 Twitter

The research topic in this master thesis is sentiment analysis on microblogging posts from Twitter. There are two possibilities to collect a data set of tweets: by using an existing data set or by crawling a completely new data set. The former is the more easy solution, but this can only be done when good data sets are available. Luckily, two useful data sets were found. They will be described in section 5.2.1. Section 5.2.2 explains how the tweets from the data sets are crawled from Twitter.

5.2.1 Existing data collections

Twitter sentiment corpus

The first data set[1] of tweets was created by Sanders Analytics, a Seattle-based startup[2]. It consists of 5513 hand-classified tweets, which makes it interesting for supervised learning. All tweets in the data set are classified as positive, negative, neutral or irrelevant.

The tweets are talking about 4 companies: Google, Apple, Microsoft and Twitter. Only the three first companies are useful, since Twitter was not on the stock market during the time interval of the tweets.

The tweets of Google, Apple and Microsoft date from October 2011. For Microsoft they range over 2 days, for Google over 1 day and for Apple over 4 days.

Twitter census: stock tweets

The second data set[3] was created by Infochimps[4], a startup that sells a platform for big data query and processing. Last year it was acquired by CSC. The data set consists 2.3 million stock tweets.

Financial tweets are characterized by the usage of the dollar sign ($). Twitter users have the convention to use $[ticker] or $$ to indicate that they will discuss stock quotes, companies or other financial oriented information.

The tweets range over the period from March 2006 until March 2010.

5.2.2 Crawling system

API limitations

Twitter made some limitations to its API users. Three of them had an impact on the data collection of this master thesis and will be explained here.

[1] http://www.sananalytics.com/lab/twitter-sentiment/

[2] www.sananalytics.com

[3] http://www.infochimps.com/datasets/twitter-census-stock-tweets/

[4] www.infochimps.com

The first limitation puts a limit on the redistribution of data. It says that if you collect a set of tweets, you are not allowed to redistribute the full tweets, only the IDs of tweets and users [57]. This implicates that the two existing data sets from section 5.2.1 consist of tweet IDs alone and not of the full tweet information.

The second limitation involves the REST API. This is one of the two APIs offered by Twitter. Applications can make requests to the REST API and Twitter will forward an answer to their requests. The alternative is the streaming API, which can be used to get streams of tweets as they occur. The REST API is rate limited, which means that applications can only do 180 API calls each 15 minutes [58].

The last limitation that affected this thesis is a limitation on search queries. It is not possible to collect a set of tweets older than 6-9 days based on a search query [61]. The only possibility of getting older tweets is by scanning full timelines of tweets, but then you need to know already which timelines offer interesting tweets.

Implementation

First, a new application was registered at the Twitter developers website to get access tokens for the Twitter API. Then, a small Java system was designed to obtain tweets. It starts from a tweet ID and then gets the full tweet information from Twitter, including the information of the user who wrote the tweet. All fields of the data structure from section 5.1.2 are filled.

The system was written in Java and uses Twitter4J[5], an unofficial Java library for Twitter. It stores all collected information in the MySQL database that was introduced in section 5.1.1.

The system takes care of the rate limit as introduced in the previous section. It makes 180 calls to the REST API each 15 minutes. It regularly checks the remaining calls and intervenes when needed.

5.3 Stock information

As said before, sentiment analysis is the research topic of this thesis. But the sentiment found in tweets of a specific company is expected to be linked to the stock quotes of that company. That is why a good data set of stock information is also necessary.

The data set needs to contain intraday data, which means that it must be fine-grained data from minute to minute. Fine-grained information is necessary because Twitter is a volatile medium. In most cases only interday data, which goes from day to day, is freely available. Two companies gave the necessary data to start the research, which are described in section 5.3.1.

[5] http://twitter4j.org/en/index.html

5.3.1 Existing data collections

Dow Jones Index

The first data collection was a gift from Price-data.com[6] and contains the intraday data of all Dow Jones stocks. It contains opening, closing, highest and lowest price and the traded volume on a minute by minute base. The period of the stock quotes goes from August 1997 until November 2013.

Apple and Google

The second data set was much smaller and contains only the intraday data of Apple and Google for the period of the tweets of the data set of Sanders Analytics (see section 5.2.1). It was obtained from Tick Data[7]. It contains again opening, closing, highest and lowest price and the traded volume on a minute by minute base.

5.4 Conclusion

This master thesis combines sentiment analysis and stock information. That is why two sorts of data sets are necessary: first a good collection of microblogging posts from Twitter is needed and secondly fine-grained stock information is needed.

The microblogging posts were collected in two distinct data sets: one from Sanders Analytics and one from Infochimps. The first data set is manually tagged with a sentiment, which makes it interesting for supervised learning.

Stock data was obtained from two companies: Price-data.com and Tick Data. They cover all stock data from the stocks of the Dow Jones Index and all stocks of the companies from the data set from Sanders Analytics during the period of that data set.

All data was stored in a MySQL database, which is managed by SQL. The database consists of different tables, representing each one a well defined concept. There is one table for tweets, one for users, one for companies, two for the data sources of tweets and one for each company with stock information.

[6] www.price-data.com

[7] www.tickdata.com

Chapter 6

Experiments

Literature shows that correlations exist between the collective mood of people, measured from tweets, and stock market indices. Most of the papers use stock indices for their research and end-of-day data. This master thesis tries to find correlations on a much smaller time interval and uses stock data with a granularity of one minute. Furthermore it does not focus on stock indices of different companies, but it focusses on the stock data of individual companies.

Within papers that describe prediction systems, there is always an evolution of refinement. In the beginning, all systems start with predicting indices using end-of-day data and then further refine to individual companies using intraday data. Not much research is done by papers that use Twitter as textual source and that try to predict the stock market using fine-grained data on the company level.

The purpose of the different experiments presented in this chapter is to get a better understanding on the influence of Twitter on the stock market. The experiments test if a window of influence exists in this context, and if it is the same for experiments using tweets as the aforementioned window of 20 minutes for experiments using news articles.

This chapter is structured as follows: section 6.1 explains first how the data is preprocessed for the different experiments. Then section 6.2 describes experiments that group similar tweets together before calculating the sentiment in the tweets and linking this sentiment to the stock market. Section 6.3 tries to find answers why the stock market is sometimes correlated positively and sometimes negatively. This analysis is done by using topic modelling. The last experiments, described in section 6.4, use the Stanford sentiment annotator to determine a sentiment graph. The advantage is that this approach can also be done at unlabelled data. The chapter ends with a conclusion.

6.1 Data preprocessing

The previous chapter described how the data of this master thesis was collected. Twitter users have their own language, acronyms and habits. This is why the raw tweet data is not always direct usable for the sentiment analysis task. The tweets need first to be preprocessed. This section describes how the preprocessing is done in this thesis. The preprocessing is sequential, but it is always possible to skip one or more steps for certain experiments.

The preprocessing starts by replacing all emoticons by a generic emoticon replacer. For instance all variants of a smiley or happy face are replaced by "happy_face". Wikipedia contains an extensive list of emoticons and their meaning [65]. This list is used in this work.

The second step in the process is to convert all tweets to lower case. This is important because some users use upper case letters and some don't. This preprocessing rule assures that words are considered equal despite the use of capital letters.

Twitter users use hashtags as a form of metadata tag. Hashtags consist of a word prefixed with the number sign (#). Hashtags contain often valuable information for the understanding of a tweet. This is why hashtags are not removed, only the number sign is removed.

Stop words are not useful as indicators of sentiment. Therefore they are removed from tweets. A list of stop words [1] was obtained from A Norm Al[1], a blog from the software developer Armand Brahaj.

Times indicated by am, a.m., pm or p.m. are removed from tweets and replaced by a generic word "time". Numbers are also replaced by a generic alternative, the word "number". The use of times and numbers might be important for the sentiment, but the specific time or number is not important.

Twitter users want sometimes to put an extra stress on certain words by using repeating letters. This introduces a lot of variants of the same word. Therefore more than two letters are replaced by two letters.

Twitter users use the dollar sign ($) to indicate that they are talking about a public traded company. More specific are dollar signs followed by the ticker of the company. When they are talking about Google, they write for instance $GOOG. Tickers can be important indicators of stock market sentiment, so they should not be removed from the tweet. To make sure that $GOOG and GOOG are both considered as the same ticker, all occurrences of dollar signs are removed.

Tweets contain regularly links to other web pages. Those web pages might contain important information about the feelings of the author of the tweet; nevertheless, they are not considered in this master thesis. Therefore all URLs are replaced by the generic word "url".

Twitter users often reply on other user's tweets or address their tweets to other users. Therefore they use the at sign (@) followed by the username of the other user. All occurrences of @ followed by a username are replaced by the generic word "username".

Twitter users are frequently using marks. Those marks are removed. This preprocessing rule might be skipped or limited when it is important to understand the different parts of a sentence.

[1] www.norm.al

6.2 Clustering approaches

6.2.1 Description

The first experiment tries to find a correlation between the tweets and the reaction of the stock market. A clustering approach is followed to achieve this target. Tweets are first clustered according to their content and then compared to the evolution of the stock market. The tweets of the data set of Sanders Analytics (see section 5.2.1) are used in the different experiments, more precisely the tweets of Apple and Microsoft. The rest of the tweets could not be used here because they are from a period where the stock was not open (for Google) or a period where the company was not public in the stock market (for Twitter). Table 6.1 gives the distribution of positive, negative and neutral companies in the different companies. The number of neutral tweets is much higher than the number of positive or negative tweets.

Company	Number of tweets			Total
	Positive	Negative	Neutral	
Apple	150	301	495	946
Google	193	49	560	802
Microsoft	89	124	609	822
Twitter	52	60	532	644
Total	484	534	2196	3114

TABLE 6.1: Distribution of tweets in the data set of Sanders Analytics

K-means clustering is a popular method for clustering analysis. It was introduced in 1967 by James MacQueen [32]. The method aims to divide observations in k clusters. It consists of an initial phase followed by two iterative steps. In the initial phase k initial means are randomly generated. Then, in the first iterative step, all observations are assigned to one of the k clusters according to their distance to the mean. The second iterative step calculates the centroids of the clusters; they become the new means of the clusters. Finally those two iterative steps are repeated until convergence is reached. The algorithm could be seen as a variant of the expectation-maximization algorithm, which was described in section 3.3.3.

Several parameters are tested in the different experiments. The first is the clustering method. A baseline is formed by clustering the tweets according to the manually annotated tag. This results in three clusters (positive, negative and neutral). Hereafter clustering based on the k-means algorithm is performed, with varying size of k.

The second parameter is the time lag between the appearance of new tweets and the reaction of the stock market. This time lag varies: 1, 5, 10 or 20 minutes. The minimum interval is 1 minute, identical to the granularity of the stock market data. A 20 minute interval is taken as upper limit, because several papers from the literature describe that a window of influence of 20 minutes is the most interesting for stock market prediction.

A third parameter is the length of the interval in which tweets are grouped. During the experiments, intervals of 1, 5, 10 and 20 minutes are tested. The reasons for those interval lengths are the same as for the time lag described in the previous paragraph.

The correlation between the stock market graph and the sentiment graph was measured using the linear correlation function of Matlab [34]. This function allows measuring the correlation between two vectors and also returns how significant the result is.

6.2.2 Results

Clustering based on annotated labels

The first experiment clusters the tweets based on the annotated labels. The sentiment graph is made based on the appearance of positive, negative and neutral tweets in time. If multiple tweets occur in the same time interval, an average is calculated. Positive tweets have a sentiment score of *+1*, negative of *-1* and neutral of *0*.

Tables 6.2 and 6.3 show the correlations between the stock market graph and the sentiment graph. The results show that there are some small significant correlations, but most experiments show nothing significant. The results show no uniform positive or negative correlation. It is not possible to determine a best interval for grouping the tweets or to determine the best time lag.

Tweets interval	Time lag			
	20 minutes	10 minutes	5 minutes	1 minute
20 minutes	0.3461*	-	-	-
10 minutes	0.1184	0.1957	-	-
5 minutes	-0.0420	0.0412	0.0844	-
1 minute	-0.0516	-0.0463	-0.0436	-0.0034
$*$: $p{-}value{<}0.05$				

Table 6.2: Correlation between annotated tweets and the stock market for Apple

Tweets interval	Time lag			
	20 minutes	10 minutes	5 minutes	1 minute
20 minutes	0.0547	-	-	-
10 minutes	0.0355	-0.1589	-	-
5 minutes	0.0601	0.0437	-0.1265	-
1 minute	0.0582	0.0485	0.0155	-0.0959
$*$: $p{-}value{<}0.05$				

Table 6.3: Correlation between annotated tweets and the stock market for Microsoft

Figures 6.1 and 6.2 show two plots of the experiments using a time lag of 10 minutes and grouping tweets with intervals of 10 minutes. These intervals are not resulting in the best results for this experiment, but they are for later experiments. In this way the plots of this experiment can be compared to later plots. However, even if the results are not significant on the complete time period, some correlations (positive and negative) for small parts of the day can be seen on the graphs. Around noon on the second day of Apple a dip is followed by a peak for both the sentiment and stock graph.

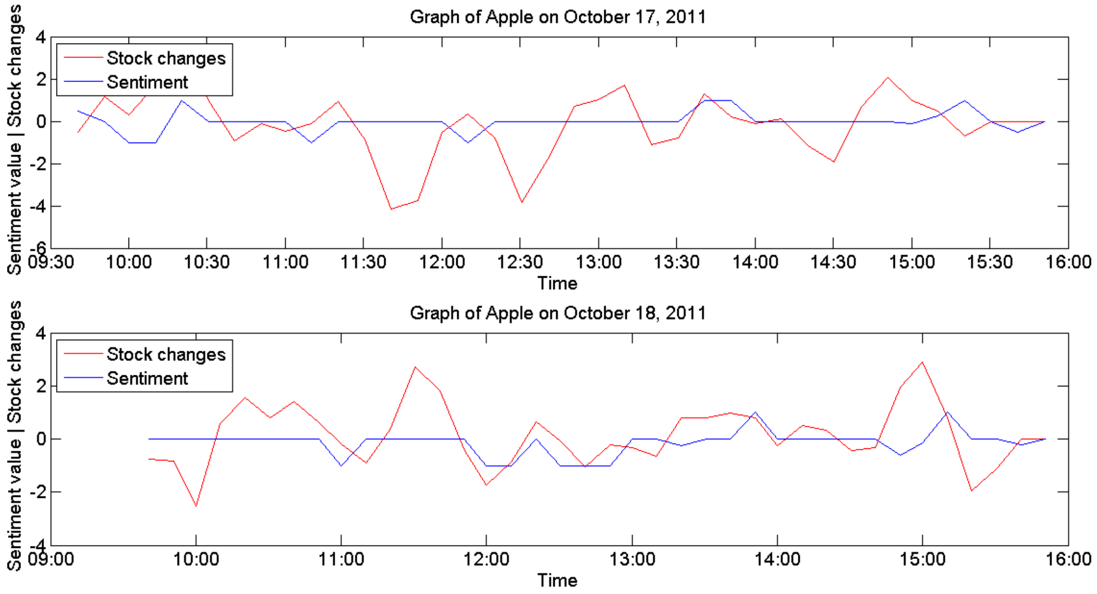


FIGURE 6.1: Graph of Apple on October 17-18, 2011

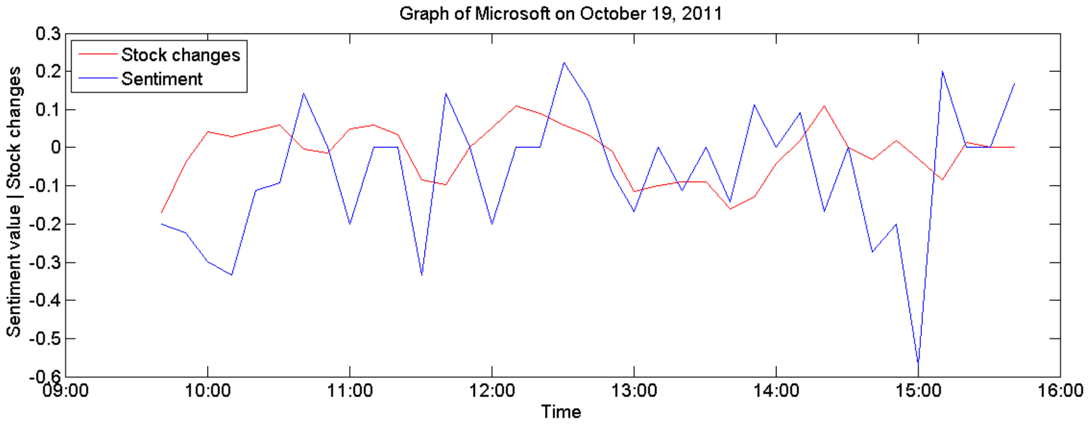


FIGURE 6.2: Graph of Microsoft on October 19, 2011

Clustering based on the bag-of-words representation

The second experiment clusters the tweets according to their content. It uses a bag-of-words representation of the preprocessed tweets. The preprocessing follows the procedure described in section 6.1. The clustering is performed by using the Matlab function for *k*-means [33].

The centroid of each cluster is calculated, including a sentiment score depending on the tweets in the cluster. The sentiment score will vary between *+1* (a cluster with only positive examples) and *-1* (a cluster with only negative examples). The sentiment graph is again calculated according to the appearance of tweets. For each tweet the sentiment score of the centroid is taken and not the manually annotated tag from the data set.

The different experiments test again several parameters: the clustering size, the length of the intervals in which tweets are grouped and the time lag between tweets and the reaction of the stock market.

Tables 6.4 and 6.5 (similar to tables A.1 and A.2) show the results of the experiments for three clusters. The other results are presented in section A.1 of appendix A. We can see that the number of clusters is not critical for the results, the results are comparable over the different experiments.

Tweets interval	Time lag			
	20 minutes	10 minutes	5 minutes	1 minute
20 minutes	0.3836	-	-	-
10 minutes	0.3470*	0.3559*	-	-
5 minutes	0.0648	0.1408	0.0415	-
1 minute	0.0453	0.0702	0.0736	-0.0944
$*$: $p{-}value{<}0.05$				

Table 6.4: Correlation between clustered tweets and the stock market for Microsoft for 3 clusters

Tweets interval	Time lag			
	20 minutes	10 minutes	5 minutes	1 minute
20 minutes	-0.2418	-	-	-
10 minutes	-0.2408*	-0.2548*	-	-
5 minutes	-0.1253	-0.1507	-0.1137	-
1 minute	-0.0937*	-0.1200*	-0.0965*	-0.0261
$*$: $p{-}value{<}0.05$				

Table 6.5: Correlation between clustered tweets and the stock market for Apple for 3 clusters

In contrast with the experiments with the annotated data, some clear trends appear in the correlations now. The tweets of Microsoft are mostly positively correlated to the stock market, whereas those of Apple are negatively correlated. It is also clear that a time lag of 10 minutes gives the most significant results. Furthermore the results show that if the tweets are grouped in intervals of 10 minutes, the correlations are the most significant.

Table 6.6 summarizes the results for experiments performing experiments with a time lag of 10 minutes and tweets grouped in intervals of 10 minutes. The tables show that 5 results are significant, while two others have a p-value higher than the threshold of 0.05 but lower than 0.1. Figures 6.3 and 6.4 (similar to figures A.7 and A.8) show the graphs of Apple and Google for those experiments, section A.2 of appendix B shows the plots for other clustering sizes.

Number of clusters	Apple	Microsoft
3	-0.2548*	0.3559*
5	-0.2392*	0.2862^1
7	-0.2318*	0.2380
9	-0.1917^2	0.3247*
$*$:$p-value$<0.05 ^{1}:$p-value$=0.0859 ^{2}:$p-value$=0.0972		

Table 6.6: Summary of correlations for tweets grouped in intervals of 10 minutes and time lags of 10 minutes

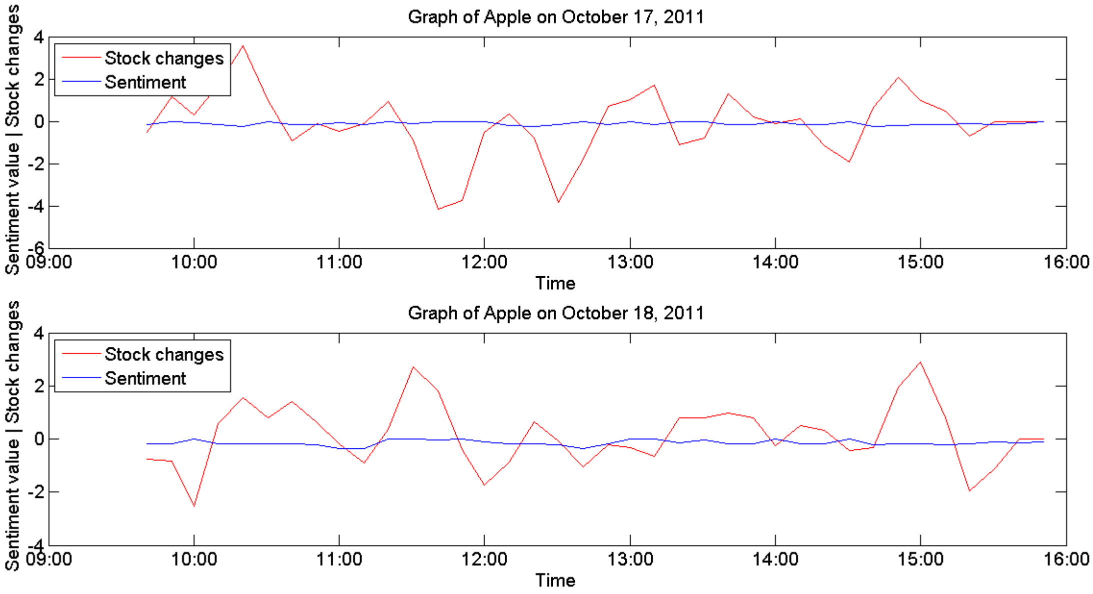


Figure 6.3: Graph of Apple on October 17-18, 2011

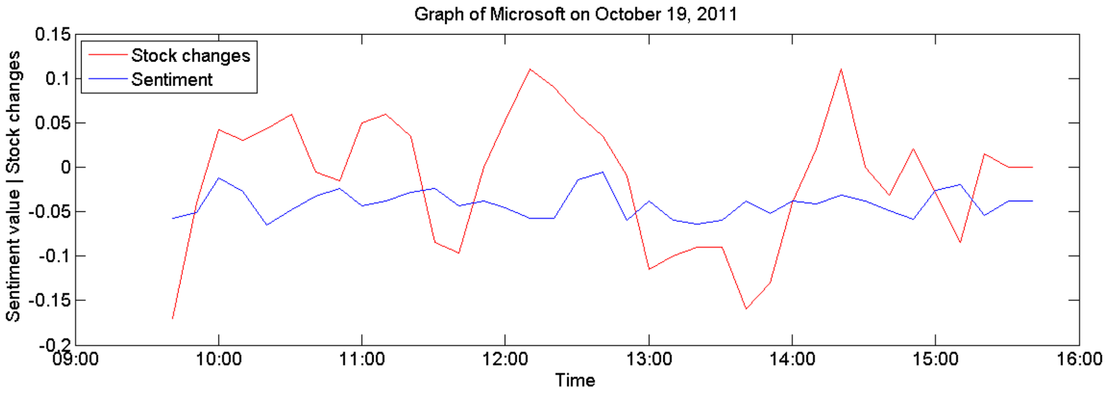


Figure 6.4: Graph of Microsoft on October 19, 2011

6.2.3 Conclusion

The experiments show that the clustering approach returns more consistent results than the annotated approach: for Microsoft a positive correlation and for Apple a negative correlation. A combination of a tweets grouping interval and a time lag of both 10 minutes performs the best.

Even if at first sight the results seem promising, more research is needed, because there are several unanswered questions. Why is the correlation for one company positive and for the other negative? Also, only two companies were tested in a specific time period, which is not sufficient to draw general conclusions from.

More experiments will be needed to get more insights in the data and the relations between tweets and stock market.

6.3 Topic models

6.3.1 Description

The second experiment has as goal to get a better understanding of the data set of Sanders Analytics in order to get better insights in why the Twitter data is sometimes correlated positively and sometimes negatively with the stock market changes. This experiment tries to find out which topics characterize the different manually annotated sentiments: positive, negative and neutral. Furthermore it also tries to look deeper into those topics to see which words have the most influence in the different models, and which words are repeated across the different companies of the data set. The complete data set of Sanders Analytics is used, that means tweets of four companies.

For that purpose, in this thesis latent Dirichlet allocation (LDA) was used. LDA was introduced in 2003 by Blei et al. [4]. It was proposed as a generative probabilistic model for topic modelling. Hong and Davison showed in 2010 that it could also be used in the context of Twitter [25]. It is a useful method to answer the research questions of this experiment.

The system in this thesis uses the Matlab Topic Modelling Toolbox of Steyvers and Griffiths [23]. The toolbox needs three input vectors to build an LDA model. The first is a vector containing the vocabulary. The other two form together a bag-of-words representation of the different documents. The tweets are preprocessed for this bag-of-words representation as earlier described in section 6.1.

6.3.2 Results

An LDA model with 20 topics was made of all tweets of each sentiment, for all companies together and for each company apart. This gives the opportunity to study the different topics that are discussed in the positive, negative and neutral tagged tweets of the different companies and gives insights in the words used by Twitter users.

Table 6.7 shows the first five topics of each category (positive, negative and neutral) for the LDA on all tweets together. The results of the other LDA topics can be found in appendix B.

	Topic number	Most important words
positive	1	ios, better, mango, right, upgrade, down, guy, info, live
	2	siri, it, technology, actually, add, thx, happy, support, things
	3	shows, cant, look, make, new, feel, little, blown, thats
	4	android, google, ics, sandwich, ice, galaxynexus, data, looking, doing, idea
	5	rt, users, people, service, facebook, ive, wait, way, didnt
negative	6	microsoft, apple, fail, windows, hate, day, steve, eclipsed, skype, people
	7	issues, need, thanks, make, hold, software, buy, hell, youre
	8	ios, problem, not, accounts, report, thats, think, wont, annoyed
	9	store, service, customer, genius, apps, life, sucks, w, using
	10	new, fucking, work, app, rt, ipad, like, hey, fuck
neutral	11	little, newly, xbla, coming, technological, announces, onl, kinda, learned
	12	twits, im, knololi, followers, neea, yoo, wannabe, leopard, autopilot
	13	google, searay, got, mobil, digital, multitouch, bookcase, funny, possibly
	14	ipads, apple, sinking, dont, big, watson, hey, dear, used
	15	going, lovatics, talent, consider, iphoneartists, let, more, beat, bus

Table 6.7: First five topics of each category of the LDA on all tweets

From the results of the LDA, it becomes immediately clear that many tweets contain customer opinions and reviews on old and new products of the different companies. For instance, different mobile operating systems appear in the topics of table 6.7: iOS (Apple), Mango (Microsoft) and Android Ice Cream Sandwich (Google). This is in line with the expectations: Twitter is a popular medium for people to share thoughts and opinions.

Certain terms appear multiple times in the topics, in both the positive and negative topics. This is due to the different experiences of the Twitter users with the products described by those terms. The word "update" appears for instance in all three classes of Apple (see tables B.4, B.5 and B.6) . The positive appearance associates words as "appreciated" and "loving" to "update", whereas the negative associates "pissed", "fix" and "annoyed" to it. A suggested interpretation of this phenomenon can be that immediately after the announcement of new features people seem to be enthusiastic, writing positive tweets. When they have tried the new features or products, this enthusiasm changes in criticism and then they write negative tweets. Neutral tweets just describe that there are new features without going more into detail.

As explained before, the results of the LDA show that many topics contain names of products and services of the different companies. These names are not only company specific, but also time specific. If you do the same LDA on another time period, other topics might appear due to product upgrades or renames. Still it is possible to associate the found topics to sentiments, because several words within the different topics announce the sentiment. When you look to the table 6.7, you see several positive words appearing in the different topics. The same is true for the negative tweets. These positive and negative words can be detected in an automatic way by using tools as SentiWordNet. This tool was earlier described in the context of Twitter and stock market prediction in the work of Chen and Lazer, see section 4.3.3.

6.3.3 Conclusion

Twitter users tweet about the products that companies announce and about the products that they have tested. The tweeted opinions and reviews contain different indicators of sentiment, but their value for the prediction of the stock market might be lower than hoped for. The expectations of the Twitter users on a certain moment in time do not necessary correspond with the same expectations of traders on the stock market.

6.4 Stanford sentiment annotator

6.4.1 Description

The previous experiments showed that small correlations between tweets and the stock market occur. Small words provide the sentiment of a tweet. Nevertheless, even if those results are promising, they are not tested on unlabelled data. Therefore this last experiment uses the Stanford sentiment treebank, introduced by Socher et al. [55], to determine the sentiment of a sentence based on its structure.

In this experiment the results of earlier experiments are followed: the tweets are grouped in intervals of 10 minutes and a time lag of 10 minutes is used. The Stanford annotator determines the sentiment of the tweets, followed by grouping them together to form the sentiment graph.

The big advantage of this experiment is that it does not require human intervention. Both data sets described in section 5.2.1 can be used for this experiment.

6.4.2 Results

Three experiments will be described here. The first two involve the data set of Sanders Analytics, because it allows to compare the results of this section to those of section 6.2. The third experiment was done on the tweet and stock data of October 22, 2009 for Microsoft. On this date Microsoft released Windows 7. This resulted in the day with the most tweets of all Dow Jones stocks of the data set described in section 5.2.1.

The first two experiments, using the data of Sanders Analytics, show for both Apple and Microsoft a slightly negative correlation. For Apple one of -0.0139 and for Microsoft -0.1379. Both results have a p-value higher than 0.1 and cannot be considered as significant.

Figures 6.5 and 6.6 show the plots of these experiments. The plots show that some correlations seem to exist during some parts of the day; for instance for the second day of the plot of Apple from 10:30 until 13:00 there are some trends that occur on both the stock and sentiment graph. This result confirms earlier results described in section 6.2.2.

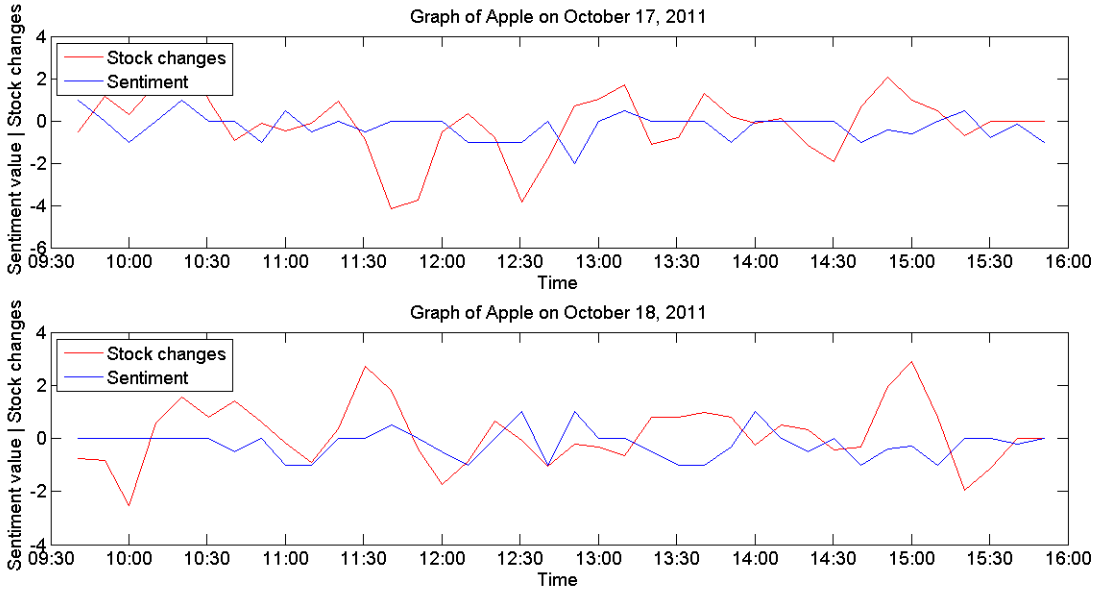


FIGURE 6.5: Graph of Apple on October 17-18,2011

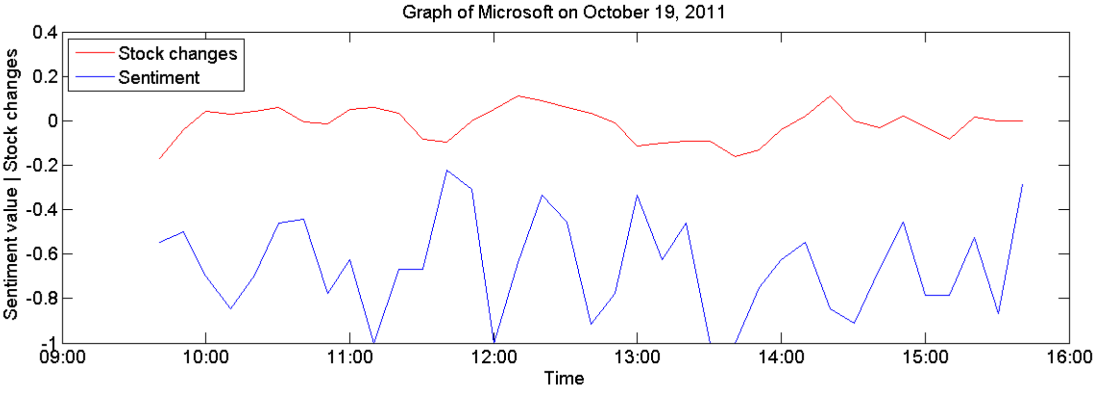


Figure 6.6: Graph of Microsoft on October 19,2011

The third experiment, using the data of Microsoft during the Windows 7 release, gave a correlation of 0.0910. This result was also not significant. The graphs of this experiment are plotted in figure 6.7 and show again some promising results for smaller parts of the day. The first 10 intervals of 10 minutes are correlated by 0.6420, while also a correlation of -0.5247 was found for 10 consecutive intervals starting at 11:00. Both results are significant.

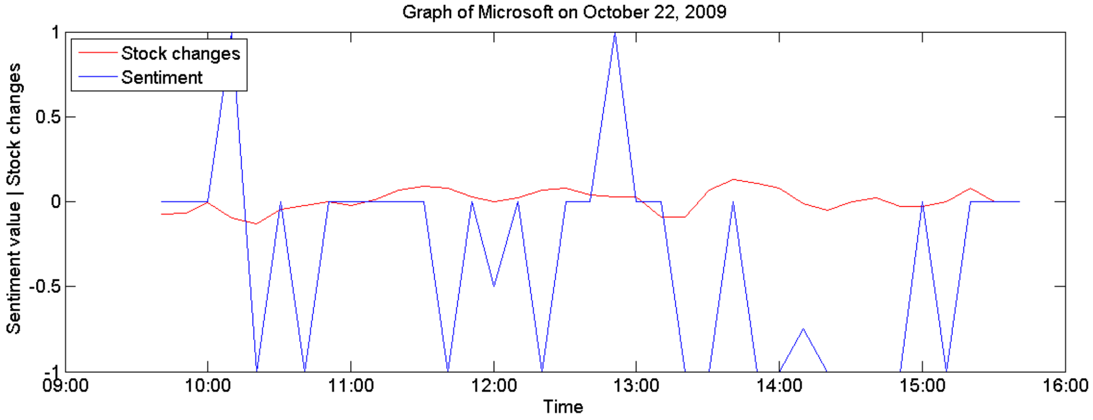


Figure 6.7: Graph of Microsoft on October 22, 2009

6.4.3 Conclusion

From the experiments with the data used in this thesis, a correlation between tweets and stock market for a whole day was not found. This implies that it is not possible to make conclusions that hold for all companies. We can see that for some moments in the day, the correlation becomes significant for multiple consecutive intervals. This suggests that Twitter can sometimes help as a stock indicator. Nevertheless, to what extend is still unclear. More research is needed to find the exact reasons.

One possible explanation would be that Twitter is only one of the hundreds of factors that influence the stock market. Its influence varies in time, which declares that the overall significance is often low, but that there are parts of the day in which it is high. For stock investors, it is important to detect these moments as soon as possible.

6.5 Conclusion

In this chapter several experiments are described that tried to find out if tweets influence the stock market. The experiments looked for correlations between the tweets and the stock market on the one hand, and the reasons behind those correlations on the other hand.

Multiple ways were proposed for finding the sentiment of tweets. The first was comparing the tweets to other similar tweets by a clustering approach. A second possibility is looking for words that indicate the sentiment. The results of the LDA topics showed that certain words within the topics can help to determine the sentiment. A third possibility is using the Stanford sentiment treebank, which can determine the sentiment of a whole sentence.

The experiments with the data used in this master thesis, show that a window of influence of 10 minutes could exist. Furthermore the best results were obtained when tweets were grouped within intervals of 10 minutes.

Some significant results were found, but for many experiments the overall significance is low. Some parts of the day have a higher significance, which might suggest that the influence of Twitter on the stock market varies in time (could be in periods of minutes, hours, days or even years). What seems clear is that Twitter is only one of the hundreds of factors that influence the stock market.

Chapter 7

Conclusion

7.1 Summary of the delivered work

7.1.1 Introduction on economic principles and sentiment analysis

The first two chapters provided the necessary background knowledge that was needed to understand the rest of this master thesis.

Financial markets are markets where people and entities can trade financial assets. One type of a financial market is the stock market, which allows investors to buy and sell equity securities from publicly traded companies. Researchers do not agree upon the possibility of predicting the stock market. Some believe that it is possible, while others think that it is impossible. Fundamental analysts use economic parameters to determine the intrinsic value of a company. If this intrinsic value is different from the price of the company on the stock market, a signal to buy or sell is created. Technical analysts try to predict the stock market by looking to historical data of stocks: prices and volumes. They use different strategies to find out how the stock is moving, for instance by using chart patterns and technical indicators.

Sentiment analysis tries to find out the attitude of the writer or speaker towards the topic he is talking about. Sentiment analysis tasks consist typically of three parts. The first step is finding a good data set. Thanks to the World Wide Web, this is a rather easy task. The second step is selecting sentiment features. Sentiment features are features that help to determine the sentiment. Four different categories can be distinguished: syntactic, semantic, link-based and stylistic features. The last step in the sentiment analysis tasks is to perform the actual analysis with an appropriate algorithm. Three kinds of methods were discussed: ad hoc rule-based approaches, supervised approaches and semi-supervised approaches.

7.1.2 Review of existing systems

Chapter 4 gave an overview of earlier work in the domain of stock market prediction. The earliest systems use news articles for their stock market prediction. Later other sources are added to the research, such as financial reports and microblogging posts.

The different papers do not agree on which method is the best approach to tackle the problem. The first thing where they disagree on is which sentiment features that they should use. Some believe in a bag-of-words representations, others use n-grams or POS-tags. Other papers use existing tools to determine the sentiment of texts: SentiWordNet, Google-Profile of Mood States (GPOMS) or OpinionFinder. The papers follow also different approaches to predict the stock market; they use SVMs, naive Bayes classifiers, probabilistic rules, latent semantic analysis and k-means.

Most papers claim that they found significant correlations between textual sources and the stock market, and that they could build a predicting system that makes profit. Most systems are however not making profit in a realistic simulation, because they suffer from high transaction costs. Those transaction costs are ignored in most papers, although they can become very high.

A window of influence opening 20 minutes before and closing 20 minutes after the appearance of a news article was discovered by Gidófalvi and widely adopted in other papers. The window of influence is the time period for which news articles have an impact on stock prices.

7.1.3 Design of a sentiment analysis system

Chapter 6 examined the influencing power of Twitter on stock prices. This master thesis tried to find correlations on a short time interval for individual companies and not for stock indices as many other papers do. The expectation was that the time lag between the publication of tweets and the reaction of the stock market is short because there are a lot of tweets appearing every moment.

First a good collection of tweets and stock data was collected. The stock data needed to have a granularity of one minute, because the system would search for correlations on a very short term. Two data sets of tweets were found, containing IDs of interesting tweets. A crawling system was designed to gather the full information about the tweets and the user who wrote that tweets.

Then a clustering approach was followed to determine the sentiment from the tweets. This resulted in sentiment graphs that could be linked to the stock data graphs. The best results were found when tweets are grouped together in time intervals of 10 minutes and when there is a time lag of 10 minutes between the appearance of a tweet and the reaction of the stock market.

The results of this first experiment were not sufficient to make general conclusions on the influence of Twitter on the stock market. A second experiment was performed to find the different topics that characterize the manually annotated tweets. By using latent Dirichlet allocation, it was found that individual words can be good indicators of the sentiment. This implies that using a bag-of-words representation can be a good representation of textual sources.

The last experiments were done using the Stanford sentiment treebank. This allows to determine the sentiment of a sentence based on its structure. The use of it can be interesting because it does not require labelled data as input. This means that it could also perform a sentiment analysis on other data than the manually annotated data set, which is important to extend the possibilities for experiments.

The experiments showed that Twitter can have a strong temporal influence on the stock market, but that it is not influencing permanently the stock market. This is somewhat expected, there are many factors that influence a complex system as the stock market.

7.2 Evaluation of the delivered work

The different experiments were performed in a logical order and built further on the outcomes of earlier experiments. The experiments did not find a big correlation between tweets and the stock market, only some temporal correlations. This is in contrast with earlier papers that claimed that they found correlations.

The used clustering approach was not the most popular in previous work and was maybe too simple to express the complex sentiments of tweets. The manually tagged data was limited, which limits the applicability of the results of the first series of experiments to the whole stock market.

The master thesis did not try to predict the stock market, as only a temporal influence of the stock market was found. Therefore it is difficult to compare the outcome of this work to earlier work.

7.3 Suggestions for future work

The experiments in this master thesis were all performed on historical data from some years ago. It would be great to see the outcome of the experiments on a more actual data set. It was not possible to perform these experiments in this work for two reasons. The first reason was that Twitter only allows to search for keywords for a very limited period in history. A real time Twitter crawler could solve this problem. A second reason is that you need stock data for the same time period and the data set for this master thesis was limited until November, 2013.

Another improvement would be increasing the amount of annotated data so that the first series of experiments could also be performed on another period in time or for other companies. This would allow to make conclusions that hold for more companies.

Future work could also use the results the LDA experiments, which indicated that some words in tweets can help to determine the sentiment of tweets. Thesauri as SentiWordNet can help with the identification of those words.

Furthermore a prediction system could be made based on a buy or sell signals that come from the temporal correlations between Twitter messages and the stock market. When suddenly a strong positive correlation appears, the system can follow the sentiment of the tweets and invest. When a negative correlation appears, the system can sell stocks. To do this, more research is needed on when such temporal correlations start exactly.

This could be done by using a thesaurus as SentiWordNet.

Appendix A

Clustering results

A.1 Correlation tables

A.1.1 3 clusters

Tweets interval	Time lag			
	20 minutes	10 minutes	5 minutes	1 minute
20 minutes	0.3836	-	-	-
10 minutes	0.3470*	0.3559*	-	-
5 minutes	0.0648	0.1408	0.0415	-
1 minute	0.0453	0.0702	0.0736	-0.0944
$*$: $p{-}value{<}0.05$				

Table A.1: Correlation between clustered tweets and the stock market for Microsoft for 3 clusters

Tweets interval	Time lag			
	20 minutes	10 minutes	5 minutes	1 minute
20 minutes	-0.2418	-	-	-
10 minutes	-0.2408*	-0.2548*	-	-
5 minutes	-0.1253	-0.1507	-0.1137	-
1 minute	-0.0937*	-0.1200*	-0.0965*	-0.0261
$*$: $p{-}value{<}0.05$				

Table A.2: Correlation between clustered tweets and the stock market for Apple for 3 clusters

A.1.2 5 clusters

Tweets interval	Time lag			
	20 minutes	10 minutes	5 minutes	1 minute
20 minutes	0.2526	-	-	-
10 minutes	0.2443	0.2862	-	-
5 minutes	-0.0078	0.0845	-0.0235	-
1 minute	0.0135	0.0405	0.0432	-0.0894
$*$: $p\text{−}value$<0.05				

Table A.3: Correlation between clustered tweets and the stock market for Microsoft for 5 clusters

Tweets interval	Time lag			
	20 minutes	10 minutes	5 minutes	1 minute
20 minutes	-0.2436	-	-	-
10 minutes	-0.2325*	-0.2392*	-	-
5 minutes	-0.1229	-0.1422	-0.1137	-
1 minute	-0.0930*	-0.1152*	-0.0948*	-0.0278
$*$: $p\text{−}value$<0.05				

Table A.4: Correlation between clustered tweets and the stock market for Apple for 5 clusters

A.1.3 7 clusters

Tweets interval	Time lag			
	20 minutes	10 minutes	5 minutes	1 minute
20 minutes	0.2511	-	-	-
10 minutes	0.1884	0.2380	-	-
5 minutes	0.0711	0.1728	-0.0162	-
1 minute	0.0174	0.0482	0.0278	-0.0538
$*$: $p\text{−}value$<0.05				

Table A.5: Correlation between clustered tweets and the stock market for Microsoft for 7 clusters

Tweets interval	Time lag			
	20 minutes	10 minutes	5 minutes	1 minute
20 minutes	-0.2177	-	-	-
10 minutes	-0.2235	-0.2318*	-	-
5 minutes	-0.1180	-0.1416	-0.1135	-
1 minute	-0.0900*	-0.1134*	-0.0973*	-0.0289
$*$: $p\!-\!value\!<\!0.05$				

Table A.6: Correlation between clustered tweets and the stock market for Apple for 7 clusters

A.1.4 9 clusters

Tweets interval	Time lag			
	20 minutes	10 minutes	5 minutes	1 minute
20 minutes	0.2353	-	-	-
10 minutes	0.1575	0.3247*	-	-
5 minutes	0.0179	0.1972	0.0376	-
1 minute	-0.0340	0.0180	0.0195	-0.0594
$*$: $p\!-\!value\!<\!0.05$				

Table A.7: Correlation between clustered tweets and the stock market for Microsoft for 9 clusters

Tweets interval	Time lag			
	20 minutes	10 minutes	5 minutes	1 minute
20 minutes	-0.1966	-	-	-
10 minutes	-0.1748	-0.1917	-	-
5 minutes	-0.1301	-0.1355	-0.1208	-
1 minute	-0.0838*	-0.0987*	-0.0726	-0.0174
$*$: $p\!-\!value\!<\!0.05$				

Table A.8: Correlation between clustered tweets and the stock market for Apple for 9 clusters

A.2 Correlation graphs

A.2.1 3 clusters

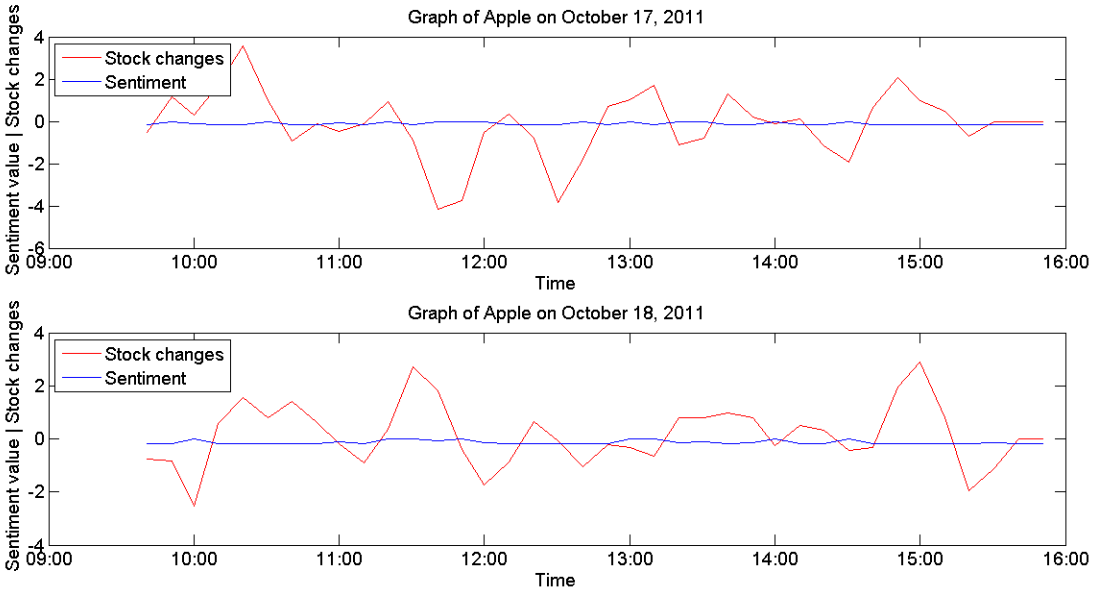


Figure A.1: Graph of Apple on October 17-18, 2011

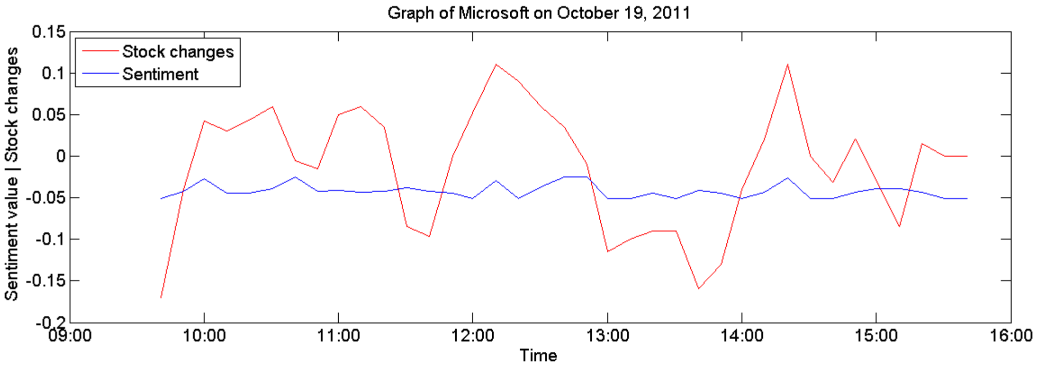


Figure A.2: Graph of Microsoft on October 19, 2011

A.2.2 5 clusters

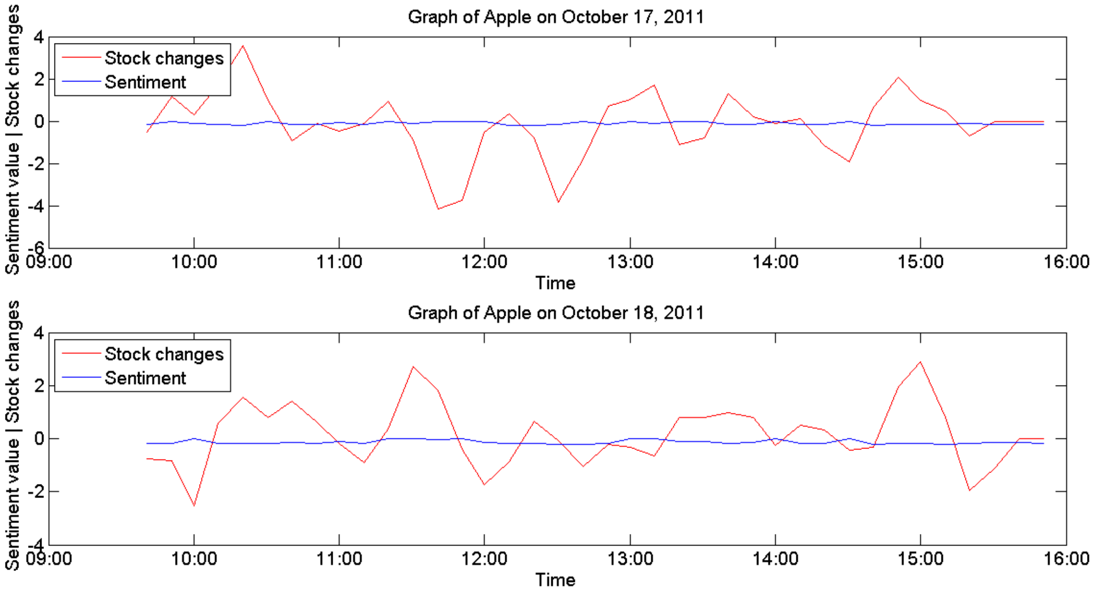


Figure A.3: Graph of Apple on October 17-18, 2011

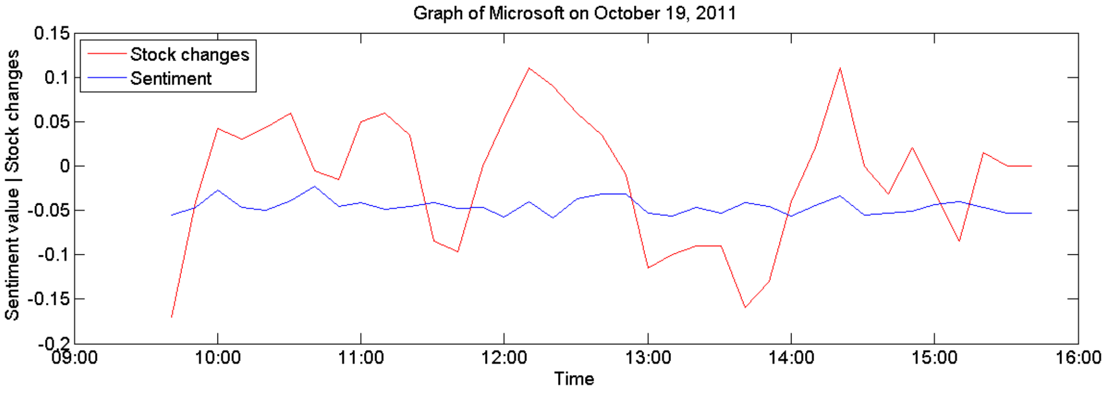
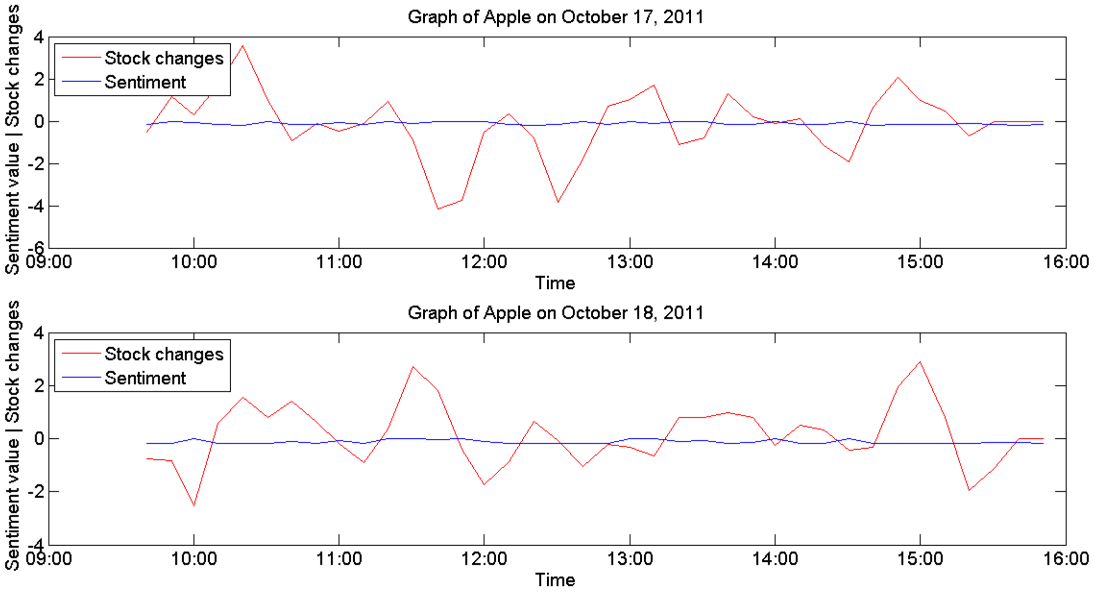


Figure A.4: Graph of Microsoft on October 19, 2011

A.2.3 7 clusters


Graph of Apple on October 17, 2011

Graph of Apple on October 18, 2011

Figure A.5: Graph of Apple on October 17-18, 2011

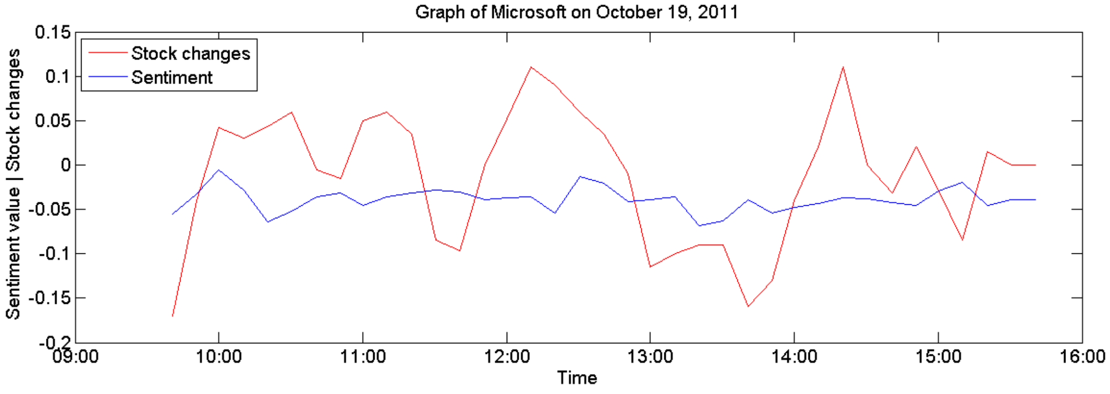

Graph of Microsoft on October 19, 2011

Figure A.6: Graph of Microsoft on October 19, 2011

A.2.4 9 clusters

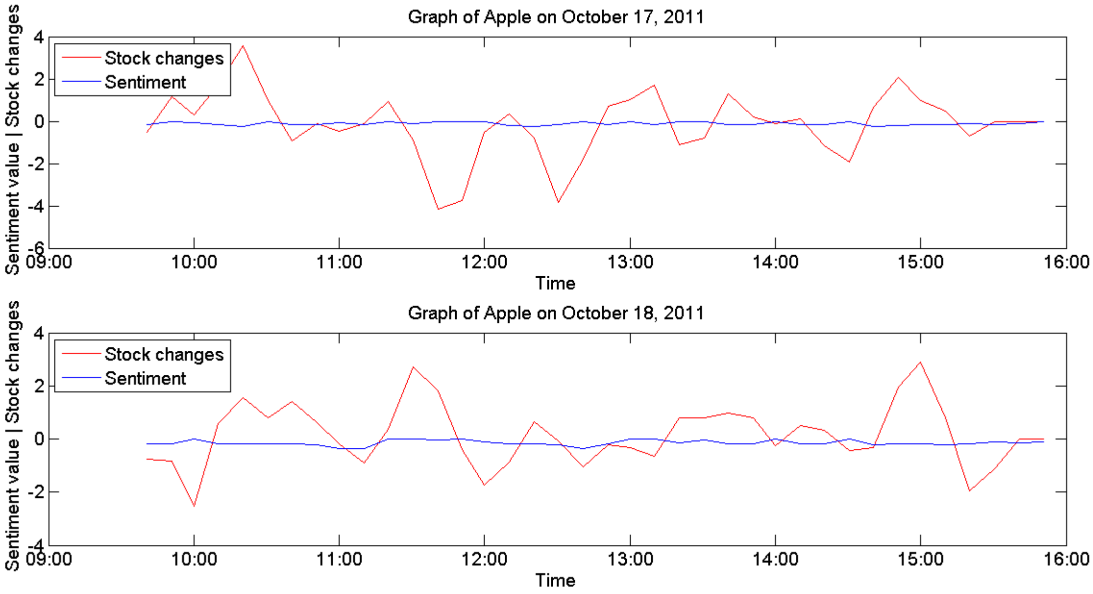


FIGURE A.7: Graph of Apple on October 17-18, 2011

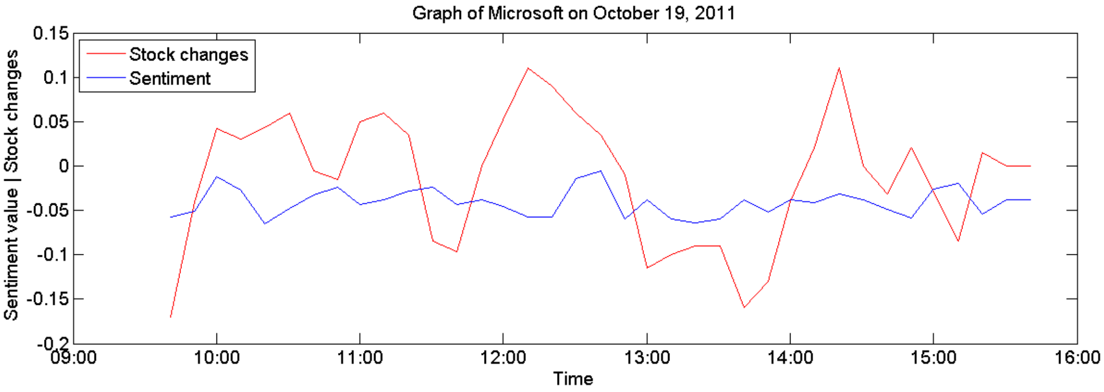


FIGURE A.8: Graph of Microsoft on October 19, 2011

Appendix B

LDA results

B.1 All tweets

B.1.1 Positive sentiment

Topic number	Most important words
1	ios, better, mango, right, upgrade, down, guy, info, live
2	siri, it, technology, actually, add, thx, happy, support, things
3	shows, cant, look, make, new, feel, little, blown, thats
4	android, google, ics, sandwich, ice, galaxynexus, data, looking, doing, idea
5	rt, users, people, service, facebook, ive, wait, way, didnt
6	os, so, start, apps, chrome, iphones, need, ny, todays
7	google, nexus, new, galaxy, looks, bitlynejbye, time, usage, smart, fast
8	i, love, like, wow, new, macbook, check, shit, thing, dear
9	good, facebook, lol, me, project, contact, dev, experience, face
10	nice, best, now, tools, finally, capabilities, icloud, literally, bitching
11	im, day, phone, did, cool, video, you, infinite, innovations
12	screen, update, finally, power, making, million, design, wait, blackberry
13	awesome, store, feature, got, sweet, email, crack, eggs, prime
14	app, thank, bookcase, ever, help, new, forward, world, appreciated
15	google, really, ipad, pretty, u, ui, excited, impressed, impressive
16	want, its, think, man, going, watching, holodesk, uh, blog
17	twitter, great, just, im, job, smartphone, googles, yeah, user
18	glad, guys, biggest, moving, pull, slick, talk, battery, buy
19	microsoft, phone, free, windows, search, improvements, devices, stores, tech, away
20	iphone, thanks, apple, today, loving, days, again, hey, research

Table B.1: Results of the LDA on all positive tagged tweets

B.1.2 Negative sentiment

Topic number	Most important words
1	microsoft, apple, fail, windows, hate, day, steve, eclipsed, skype, people
2	issues, need, thanks, make, hold, software, buy, hell, youre
3	ios, problem, not, accounts, report, thats, think, wont, annoyed
4	store, service, customer, genius, apps, life, sucks, w, using
5	new, fucking, work, app, rt, ipad, like, hey, fuck
6	better, siri, ics, need, lion, future, glad, out, theres
7	just, trouble, fixed, live, mac, updates, hear, presentation, already
8	dear, wtf, know, microsofts, release, albatross, connect, freeze, gotta
9	you, tell, isnt, me, its, want, went, ballmer, sound
10	i, dont, its, good, lose, on, help, products, tv
11	twitter, rt, fix, cant, siri, rts, goes, week, ios
12	use, going, did, apparently, calls, line, mentions, adobe, fuck
13	time, icloud, upgrade, wow, blue, dead, hour, late, bird
14	phone, shit, it, apps, let, send, bad, come, failed
15	doesnt, network, working, screen, xbox, ive, able, ass, data
16	iphone, update, itunes, needs, updating, guys, reservation, stop, up
17	battery, button, cloud, oh, ppl, sure, crashing, mail, message
18	google, android, really, now, didnt, hours, recognition, like, thing
19	restore, seriously, lot, does, retweets, thank, this, r, ugh
20	im, having, wait, computer, right, facebook, u, great, product

Table B.2: Results of the LDA on all negative tagged tweets

B.1.3 Neutral sentiment

Topic number	Most important words
1	little, newly, xbla, coming, technological, announces, onl, kinda, learned
2	twits, im, knololi, followers, neea, yoo, wannabe, leopard, autopilot
3	google, searay, got, mobil, digital, multitouch, bookcase, funny, possibly
4	ipads, apple, sinking, dont, big, watson, hey, dear, used
5	going, lovatics, talent, consider, iphoneartists, let, more, beat, bus
6	microso, window, phne, cloud, free, noiva, available, touchpoints, windowslivebook, nem
7	ballmer, sterling, ya, microsoft, jobely, ceo, luck, bying, steveb
8	business, packed, words, tweeps, cool, following, shari, women, novo
9	tear, bing, days, hours, minute, yeah, second, sentences, age
10	lik, uses, soaddicted, did, related, smart, tipredirect, excel, feel
11	itrt, tim, good, runs, ready, confirms, face, gingerbread, lookout
12	google, android, nexttime, ice, galaxy, san, ics, dhilipsiva, galaxynexus, scvmm
13	rsna, twits, facebook, follow, pen, realized, best, day, nigh
14	thank, make, offers, tear, johnnyvegas, ohemgee, ma, statuses, managed
15	app, totally, workspace, hello, willprovide, iphoto, baby, googleplus, memorial
16	networkspecs, phone, user, sconds, milli, apps, solar, data, wait
17	vp, what, io, longfor, socialize, better, tomcruise, hong, knw
18	i, jus, usdor, major, dont, logo, does, isv, ip
19	great, tá, fb, thang, googles, which, launches, ouch, gone
20	stops, vic, blog, check, shared, genius, buy, buzz, steveballmer

Table B.3: Results of the LDA on all neutral tagged tweets

B.2 Apple

B.2.1 Positive sentiment

Topic number	Most important words
1	i, million, getting, people, red, apologize, beta, companies, company
2	thanks, service, education, gave, gratis, hand, v, affair, current
3	phone, nice, like, bar, using, hoping, looking, replacing, steve
4	great, time, guys, lost, pretty, tone, admint, att, beautiful
5	got, great, sweet, case, email, sent, smartest, alive
6	technology, said, bless, changed, create, dad, easy, eggs, grandma
7	iphone, store, feature, ipad, thx, upgrade, customer, twitter, easter, iphones
8	love, ios, apple, actually, apps, absolutely, backside, down, a
9	i, siri, its, help, it, air, changed, going, ipads, years
10	rt, days, job, replaced, icloud, argument, bag, body, brilliance
11	macbook, best, kind, post, pro, switch, this, truly, any
12	good, guy, best, done, eggs, feel, finally, making, publicly
13	ever, man, middle, bad, card, forward, ive, so, stevejobs
14	just, crack, brother, stewart, thank, welcome, attempts, bravo, cardsapp
15	iphone, screen, happy, ceremoniously, currently, hairline, ipod, pulls, booted
16	thank, service, support, pull, did, glad, things, want, break
17	im, today, think, way, right, app, longer, applety, blackberry
18	awesome, doing, it, battery, bitching, day, developer, fixed, genious, little
19	free, weekend, better, blackberry burn, gets, heres, album, booked
20	new, loving, update, wow, ipad, sells, appreciated, architecture, deals, far

Table B.4: Results of the LDA on the positive tagged tweets of Apple

B.2.2 Negative sentiment

Topic number	Most important words
1	icloud, u, app, lion, trouble, hell, bug, features, tech
2	iphone, life, on, stop, restore, thats, today, apparently, awful
3	cant, make, product, think, quality, restore, definitely, lies, nothing
4	doesnt, im, computer, isnt, upgrade, wow, ipod, reservation, tell
5	fix, it, send, update, gotta, pissed, buy, r, authorize
6	apple, store, service, customer, right, hate, hours, bad, hold
7	new, iphone, update, work, now, hey, wtf, release, annoyed
8	battery, phone, itunes, better, line, products, complain, down, freeze
9	having, issues, really, w, you, did, rt, launch, them
10	wont, great, phone, worst, problem, experience, hope, ok, wanna
11	day, fuck, itunes, seriously, loop, upgrading, book, combo, control
12	dont, thanks, im, software, syncing, fixed, little, mb, post
13	apps, cant, going, screen, bar, late, youre, officialy, open
14	rt, ios, wait, its, needs, sync, didnt, restore, that
15	i, fail, its, need, use, let, really, oh, updating, guys
16	fucking, iphone, minutes, missing, sucks, button, drain, folder, fuck
17	time, ipad, not, ever, theyre, connecting, folks, hard, know
18	just, dear, data, good, disappointed, ive, asked, boycotting, connect
19	siri, shit, like, network, does, apps, voice, iphones, out
20	genius, me, come, contacts, help, forever, half, properly, sure

Table B.5: Results of the LDA on the negative tagged tweets of Apple

B.2.3 Neutral sentiment

Topic number	Most important words
1	ipad, thanks, active, check, stores, things, u, doesnt, oh
2	, like, its, dear, icloud, buy, isnt, think, trying
3	im, sales, interesting, phone, buying, having, honor, ive, sure
4	vs, yet, stay, days, lot, sells, didnt, steves, balzora
5	does, consider, ok, community, mobile, tab, yes, att, case
6	iphone, apple, million, weekend, iphones, ipads, people, whats, phones, design
7	waiting, ipad, battery, told, awesome, cant, coming, more, s
8	apps, users, did, big, know, good, hold, earnings, end
9	app, ipod, world, hello, win, memories, touch, baby, available
10	store, twitter, best, speech, hey, better, love, angry, bidirectional
11	great, tech, line, imessage, mac, home, old, want, huge
12	going, support, device, change, come, canon, computer, hmm, read
13	stevejobs, it, talking, today, download, genius, know, life, people
14	siri, new, video, iphone, gets, got, setup, make, update
15	war, release, tribute, needs, slow, changes, global, time, call
16	just, ios, use, upgrade, android, say, tell, baby, continues
17	steve, jobs, dont, need, thank, id, u, changed, incredible
18	rt, store, sold, really, fan, eric, holder, spotted, music
19	genius, bar, that, battle, lost, way, work, you, account
20	page, want, day, siri, contract, photo, success, quarter, thx

Table B.6: Results of the LDA on the neutral tagged tweets of Apple

B.3 Google

B.3.1 Positive sentiment

Topic number	Most important words
1	google, ics, face, excited, smart, info, interesting, again, ahead
2	os, awesome, data, improvements, bookcase, search, available, helps, share
3	android, galaxynexus, live, love, way, bookcase, literally, systems, tools, announcement
4	i, want, add, finally, wow, device, make, sure, outdated, powered
5	google, users, bitlynejbye, power, video, getting, best, bookmarks, fantastic
6	google, rt, sandwich, slick, stop, telegraph, twitter, aka, blog, craving
7	im, app, looks, like, infinite, prime, unlock, event, exciting
8	phone, iphone, nice, new, wait, its, apple, day, check
9	nexus, galaxy, new, introducing, ui, fast, now, wow, dear, inbuilt
10	job, contact, user, experience, features, ios, oh, avtar, mind
11	looks, feature, know, dhilipsiva, god, need, released, respect, updated
12	android, ice, sandwich, think, company, effect, ever, go, imo, profile
13	raise, screen, work, globe, person, probably, sdk, shows, sorry
14	google, just, usage, cant, pretty, chrome, blown, guys, ive, searches
15	now, people, watching, carriers, content, font, launches, smart, thats
16	facebook, biggest, digital, hand, looking, right, sounds, threat, bag
17	really, great, look, better, design, google, app, delicious, play
18	googles, beautiful, smartphone, webgl, technology, apple, virtual, accompanying, aggregates
19	capabilities, devices, far, lol, shifted, totally, vs, yet, adopted
20	google, good, away, best, sharing, page, so, spell, thanks

Table B.7: Results of the LDA on the positive tagged tweets of Google

B.3.2 Negative sentiment

Topic number	Most important words
1	android, face, gonna, ice, its, analytics, apperently, cyanogenmod, hmm, knows
2	cloud, googleplus, wrong, chris, dennisritchie, froze, gmail, horrible, huge
3	ics, recognition, android, appthe, better, curious, ht, jajajajajajaj, needs, quickresponse
4	dear, galaxynexus, adobemax, bad, beat, ios, lecture, nexusprime, notes, response
5	sandwich, new, terrible, tons, august, demoed, dump, financial, headache, infofail
6	rt, need, buying, ebook, false, googles, introduce, load, people
7	i, just, adobe, awesome, battle, code, greatly, least, melts
8	facial, work, im, argument, awkward, blackouts, doing, done, doubledigit, facialunlock
9	phone, cool, accessibility, accounts, admirably, also, confused, crashing, dgoogle
10	dates, interference, poole, tcn, user, accessibility, accounts, actually, admirably
11	athens, illustrating, parker, uphill, verizon, vs, accessibility, accounts, actually
12	learn, please, got, hire, introduces, maximize, ouch, perils, project
13	google, presentation, encrypt, font, highlight, letdown, period, profile, quick, santorums
14	android, battery, coach, crash, docs, fails, hard, perform, plusone
15	google, didnt, facebook, nexus, not, cause, collection, demo, galaxy, googleapps
16	did, announcement, good, background, cant, economics, graphic, hell, phone
17	google, failed, thing, aint, chrome, didnt, goof, hate, moment, tried
18	fail, first, bout, demonstration, dont, electability, fix, fold, impressed
19	ics, unlock, laughable, results, revolution, source, sticky, talk
20	really, called, lot, problem, time, actually, burn, data, doubt, guys

Table B.8: Results of the LDA on the negative tagged tweets of Google

B.3.3 Neutral sentiment

Topic number	Most important words
1	google, roundup, apple, votes, ios, handson, gingerbread, marketers, make, oxycodone
2	ics, blog, via, carriers, docs, literally, lots, blogspot, brands
3	waking, wallet, buzz, im, n, competition, via, activity, demons
4	tim, tcn, coming, devices, presentation, good, update, build, cool
5	ics, picture, add, nexuss, phone, circles, isnt, tablet, yoo
6	google, send, bing, ballmer, xmas, check, pagerank, related, scale
7	apple, improvements, kml, io, updates, didnt, getting, laughs, nyquil
8	googleplus, socialmedia, gplus, analytics, facial, rewind, up, webgl, contacts
9	android, google, screenshot, available, seal, november, round, announced, reveal, tool
10	news, galaxy, unlocks, updated, feature, lapse, announce, promise, really, to
11	ice, san, unwanted, user, order, pricepoints, sandwichos, wordpress, official, astounding
12	google, bookcase, digital, infinite, announcement, released, begin, default, gallery
13	seal, searchengine, info, mark, question, stock, vegas, care, usage
14	peter, iteration, newest, small, face, big, great, verizons, data
15	i, google, facebook, twits, social, infographic, literally, novelty, perfect
16	googles, one, event, multitasking, nexusice, apps, encrypting, peepz, reveals
17	google, roundup, mob, so, going, snippet, launches, ready, slow, wikipedia
18	android, dhilipsiva, ics, offers, via, announces, available, screenshot, recently, needs
19	junkyard, running, confirms, does, dont, logitech, beat, tonights, adwords
20	networkspecs, galaxynexus, features, app, lightly, lol, what, calling, hong

Table B.9: Results of the LDA on the neutral tagged tweets of Google

B.4 Microsoft

B.4.1 Positive sentiment

Topic number	Most important words
1	windowsphone, explorer, finally, kinect, best, canada, courtesy, down, extent
2	windows, android, apple, blown, fact, faculty, innovations, month, yahoo, year
3	rt, research, google, thanks, bought, ctp, i, interest, bitlypcmjon
4	microsoft, i, details, vslive, deal, explains, impressive, science, yeah, agree
5	tech, computer, kids, love, watch, works, babygeeks, composite, dishing
6	offer, love, i, check, its, marketing, talk, arc, billions
7	mango, shows, great, holodesk, like, smartphone, technological, available, better, careers
8	local, taste, buffalo, certification, claws, didnt, fan, feeling, ibm
9	improvements, cool, sad, safe, tools, and, assessment, cant, check
10	new, holodeck, away, didnt, glad, i, users, wait, day
11	thanks, really, moving, time, absolutely, appsense, commands, foundation, going
12	b, hey, poised, citizenship, combine, did, earned, encourage, gem
13	good, help, project, bing, sale, science, siri, success, antimicrosoft, apple
14	microsoft, cloud, world, actionpretty, billgates, data, handsets, ics, interface, miley
15	free, phone, stores, devices, microsoftstores, again, here, know, funds, use
16	bi, boot, closer, contender, dotnet, enjoy, gates, offers, priceless
17	coming, music, arrives, body, door, editor, fiction, learning, let
18	enterprise, fellows, server, store, vancouver, ballmer, getting, has, hosting
19	microsoft, uh, evangelist, access, db, excel, excellent, fiction, futuristic, giants
20	search, awesome, dev, screen, start, looks, battle, change, dubbe

TABLE B.10: Results of the LDA on the positive tagged tweets of Microsoft

B.4.2 Negative sentiment

Topic number	Most important words
1	android, day, google, need, people, grcode, aggressive, arrogance, browsers
2	respect, scientist, victim, applewindows, bit, concluded, critical, currently, excel
3	access, cant, it, its, talk, victims, current, didnt, enemies
4	sake, employees, past, waiting, way, beginning, beta, blew, catastrophe
5	microsoft, skype, steve, fixed, just, hack, pc, played, agree, back
6	microsoft, like, report, ask, net, piracy, updates, word, abt, acct
7	windows, failed, failing, id, screened, agency, angst, app, ay
8	rt, nokia, phone, dont, executive, guys, acceptable, according, accused
9	live, lose, ceo, colour, future, install, microsoft, tech, alleged
10	microsofts, steveballmer, albatross, compares, crash, issue, join, neck, reader, ballmer
11	mac, explain, m, search, automatically, brands, did, forced, gate
12	adobe, dine, know, webpronews, ala, attacks, bigbrother, careers, come
13	apple, eclipsed, mole, sucks, already, choice, wtf, accounts, adweek, again
14	apps, gates, hate, says, you, blog, clear, hacked, cause
15	i, fail, u, deal, drop, lync, agree, alongside, billion, calls
16	ballmer, time, advertising, having, love, sure, them, attributed, biggest
17	microsoft, powerpoint, use, racketeering, broke, windows, accessibility, changes, collects, dumb
18	update, blue, didnt, format, mail, aargh, azure, console, mr
19	xbox, os, yahoo, fcucks, accounts, ad, backwards, business, chief, code
20	notresponding, office, a, antitrust, online, vs, byebye, cheap

Table B.11: Results of the LDA on the negative tagged tweets of Microsoft

B.4.3 Neutral sentiment

Topic number	Most important words
1	microsoft, search, bing, celebrity, crm, away, giving, need, iphone, skype
2	mobile, store, san, makes, software, diego, looks, card, kickoff
3	holiday, server, day, announces, credit, future, nov, cnet, create
4	live, omnitouch, released, users, enterprise, appsense, db, proves, salability
5	microsoft, available, technology, want, hand, web, year, researchers, spigot, partner
6	xbox, video, kinect, learning, playful, skype, business, digital, social
7	tech, learn, news, services, tablets, geeks, lync, product, getting
8	start, blog, exchange, good, hp, know, hyperv, emc, great
9	rt, new, turns, version, pc, office, sign, vmware, body
10	use, holodesk, pack, solutions, building, net, read, thinking, job
11	microsoft, windowsphone, free, mango, using, neowin, opens, zune, adcenter, calumo
12	microsoft, yahoo, lucky, buying, ceo, says, sometimes, youre, steveballmer, deal
13	touchscreen, i, just, azure, solution, coming, applications, booth, it
14	windows, cloud, phone, screen, devices, i, tcn, infosys, user, coming
15	excel, data, tips, dropx, charts, creating, rules, techedafrica, u
16	roslyn, compiler, developer, preview, project, open, snom, vmworld, webinar
17	ballmer, android, steve, microsofts, nokia, phones, ceo, computer, week, steveballmer
18	rt, sharepoint, summit, apps, beta, isnt, cisco, client, intervate
19	touch, check, surface, tomorrow, service, w, latest, people, research
20	microsoft, google, apple, vs, review, siri, patent, sqlserver, training

Table B.12: Results of the LDA on the neutral tagged tweets of Microsoft

B.5 Twitter

B.5.1 Positive sentiment

Topic number	Most important words
1	hashtags, haha, ill, man, touch, abandoned, aboutthatlife, add, addicted
2	shit, havent, anyday, feel, fresh, hurting, ive, me, twitterless
3	twitter, left, make, bitch, desperatehousewives, end, glad, literally, okayi, scares
4	people, progressive, real, forgettin, happy, time, yeahh, abandoned, aboutthatlife
5	did, mood, sleep, content, fun, gonna, keeps, makes, maybe, means
6	bed, you, way, addicted, life, min, naw, people, reliable
7	i, just, missed, today, abandoned, coo, farreal, going, sucks
8	app, hella, do, hooked, laying, looks, lovee, slightly, stalkerismo
9	apples, biggest, chao, dancemoms, ditch, forth, got, haven, shows
10	twitter, dear, characters, cnt, email, feeling, fingers, great
11	facebook, lol, new, aboutthatlife, bullshit, followme, gives, impressive, marks, need
12	i, like, appear, love, background, boo, bored, bug, caught, fwm
13	love, world, facebook, pretty, belive, bigdealdawson, engaging, enjoying, fast, fucking
14	twitter, media, answers, back, day, song, tv, twitterfacebook, abandoned
15	add, blackberryits, everytime, lightningonspeed, mad, pulling, used, abandoned, aboutthatlife
16	good, fuck, say, ass, entertaining, funny, leave, sad, abandoned
17	rt, i, beats, concerned, favorite, friend, guess, helps, introduced, lives
18	thing, helped, privacy, unlike, useful, weirdo, abandoned, aboutthatlife, add
19	im, really, social, fb, better, gotta, i, internet, letting, smm
20	it, change, connected, etc, excited, hell, isnt, little, me

Table B.13: Results of the LDA on the positive tagged tweets of Twitter

B.5.2 Negative sentiment

Topic number	Most important words
1	retweets, mentions, working, fat, argument, blowing, boring, fixit, funnyy, havent
2	cant, listing, midnight, okay, problems, tweet, able, account, acting
3	twitter, tweets, fix, bored, dms, giving, home, promoallday, soon
4	rts, tell, ass, hours, dezz, hit, muchh, replies, twitter
5	i, better, blue, got, know, distractions, fad, follow, hourly, itstolate
6	anything, procrastination, tt, caracteres, confusin, highscoolmemories, just, networkin, peopleand
7	bird, lil, rt, tomorrow, able, affects, subtweeting, tv, unable, batteries
8	nobueno, notified, ritenow, servers, study, this, timeline, tl, touch
9	having, app, doing, limit, lol, man, section, does, drawlin
10	whale, everytime, explaining, ipad, itd, rly, shitt, sucks, able
11	twitter, rt, to, bulletproof, compose, ihateyourightnow, numbers, tmw, able
12	need, retweeted, fuck, gonna, igiveup, is, me, panasonic, stupid
13	twitter, dead, asap, care, days, fixed, gotten, reporting, sleep, soo
14	ass, showin, capacity, hate, mothafuckin, now, shit, conversation, dead, difficult
15	account, able, allowing, freaking, hell, moving, y, acting, addicted
16	getting, messed, bitch, bloody, come, dear, doesnt, followlimitisgay, idontunderstand
17	twitter, problem, include, low, addicted, appreciated, dff, fuckin, nooww, pretty
18	im, cont, dont, people, acting, aint, goes, now, reason
19	facebook, fucking, computer, current, dis, done, dont, epicfail, garbage
20	focus, account, broken, emails, fawkin, feel, followers, fucked, fucking

Table B.14: Results of the LDA on the negative tagged tweets of Twitter

B.5.3 Neutral sentiment

Topic number	Most important words
1	twitter, goodnight, builds, buzz, charts, movement, occupy, street, whats, didnt
2	really, funny, z, worldwide, worry, fox, guy, mentions, ways
3	its, new, free, phone, android, ass, installed, try, twidroyd
4	dis, explain, happening, update, ah, davis, links, locks, moments
5	twitter, u, lol, sleeps, news, bed, tell, addicted, girl, book
6	off, wat, morning ,ppl, feel, them, doing, lot, nigga
7	twitter, dont, rt, night, time, best, fb, fuckyoumean, later, new
8	followers, autopilot, let, you, gain, business, gn, real, bout
9	love, wit, updating, app, forget, home, important, instead, sumbody
10	use, bitch, stop, welcome, lose, confusedbymytimeline, follow, hello, humantrafficking
11	got, know, need, help, make, w, music, youtube, care
12	twitter, da, fuck, needs, tweeps, cool, me, want, lets, acting
13	im, good, finally, cause, in, reformat, write, dms, e
14	think, bye, tonight, goin, ur, aint, dark, ff
15	like, follow, people, socialmedia, going, world, run, add, foundation
16	minutes, seconds, age, days, twittertime, hours, shit, year, people
17	rt, facebook, everybody, b, coming, join, account, ahh
18	say, post, tired, does, man, tweeting, apologize, bows, fans
19	tweet, day, tt, tweets, cuz, life, stuff, thought, back
20	i, just, addicting, again, feelings, levine, minute, opinion, sec

Table B.15: Results of the LDA on the neutral tagged tweets of Twitter

Appendix C

Poster

PREDICTING THE STOCK MARKET BY USING SENTIMENT ANALYSIS ON TWITTER

KU LEUVEN

FACULTEIT
INGENIEURSWETENSCHAPPEN

Master
Computer Science

Master thesis
Jeroen Ruytings

Promoters
Prof. dr.
M.-F. Moens
Prof. dr. ir.
M. Van Barel

Supervisor
J. C. Gomez

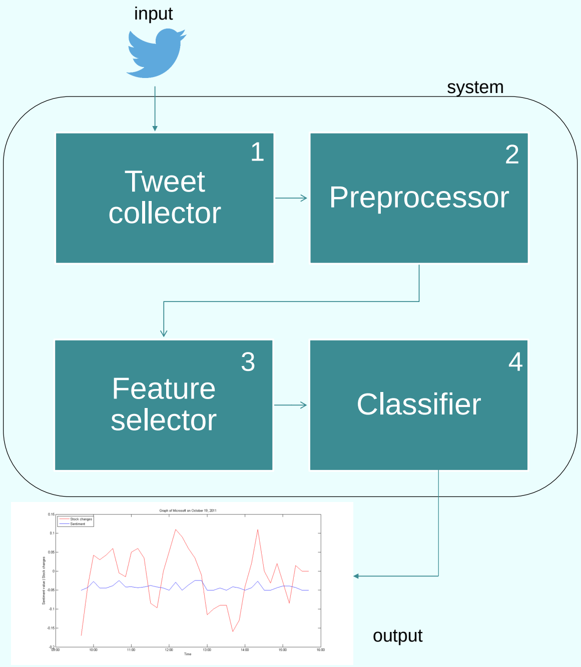
Academic year
2013-2014

Context & goals

News articles, financial reports and microblogs contain a richness of information on companies and their products. Investors would like to use that information to forecast the movements of stock prices in order to make more profit.

The stock market is a complex system, influenced by economic, political and psychological factors. This master thesis tries to find out if Twitter is also one of the factors influencing the stock market fluctuations.

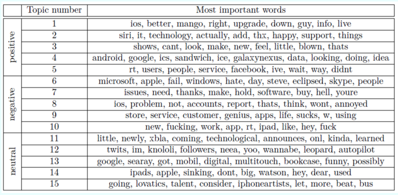
Design of the sentiment analysis system



1. Relevant documents are collected. In our system those documents are tweets containing information on companies.
2. The tweets are preprocessed to uniform features across all tweets and to remove tweet curiosities that don't contribute to the sentiment.
3. The feature selector selects the sentiment features that will be used to determine the sentiment. Our system uses a bag-of-words representation.
4. A classifier is built to determine if a text is positive, negative or neutral. This results in a sentiment graph that can be compared to the fluctuations of the stock market.

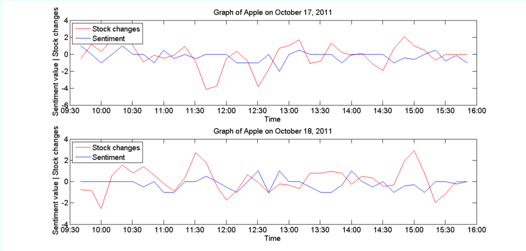
Other experiments

- Topic modelling using latent Dirichlet allocation



	Topic number	Most important words
positive	1	ios, better, mango, right, upgrade, down, guy, info, live
	2	siri, it, technology, actually, add, thx, happy, support, things
	3	shows, cant, look, make, new, feel, little, blown, thats
	4	android, google, ics, sandwich, ice, galaxynexus, data, looking, doing, idea
	5	rt, users, people, service, facebook, ive, wait, way, didnt
negative	6	microsoft, apple, fail, windows, hate, day, steve, eclipsed, skype, people
	7	issues, need, thanks, make, hold, software, buy, hell, youre
	8	ios, problem, not, accounts, report, thats, think, wont, annoyed
	9	store, service, customer, genius, apps, life, sucks, w, using
	10	new, fucking, work, app, rt, ipad, like, hey, fuck
neutral	11	little, newly, xbla, coming, technological, announces, onl, kinda, learned
	12	twits, im, knololi, followers, neea, yoo, wannabe, leopard, autopilot
	13	google, searay, got, mobil, digital, multitouch, bookcase, funny, possibly
	14	ipads, apple, sinking, dont, big, watson, hey, dear, used
	15	going, lovatics, talent, consider, iphoneartists, let, more, beat, bus

- Sentiment analysis using the Stanford sentiment treebank



Results

- Our experiments show a window of influence of 10 minutes for tweets. This is the time for which tweets might have an impact on stock prices. In literature a window of influence of 20 minutes for news articles was described.
- Determining the sentiment by grouping tweets in intervals of 10 minutes gives the best results.
- Twitter shows only temporal correlations to the stock market fluctuations. The influence of tweets on the prediction of the stock market varies during the day, months or years.

Appendix D

Paper

Het voorspellen van de beurs met behulp van sentimentanalyse op Twitter

Jeroen Ruytings
Departement computerwetenschappen
KU Leuven, België
jeroen.ruytings@student.kuleuven.be

Samenvatting

De aandelenmarkt wordt gekenmerkt door zijn complexiteit omwille van de talloze factoren die de beurskoersen beïnvloeden. De literatuur is het oneens of nieuwsberichten over bedrijven de beurs ook kunnen beïnvloeden. Vele onderzoekers beweren dat ze goede resultaten halen met een voorspellingssysteem op basis van nieuws, toch zijn maar weinig van die systemen echt winstgevend in reële situaties. Dit artikel onderzoekt de invloed van Twitter op de beurskoersen. Eerst werd er een dataverzameling aangelegd van relevante tweets. Vervolgens werd er een systeem ontworpen voor sentimentanalyse. Tot slot werd er gekeken of het geheel aan sentimenten uit de Twitterberichten een invloed hadden op de beurs. De resultaten toonden geen permanente verbanden tussen de tweets en beurskoersen. Ze toonden wel aan dat Twitter een tijdelijke invloed op de beurs kan hebben. Dit zou betekenen dat Twitter één van de factoren is die de beurs beïnvloedt, maar dat deze invloed varieert in de tijd.

1. Introductie

Het kunnen voorspellen van de aandelenmarkt is de grote droom van elke investeerder: het zou hem toelaten om gericht beurstransacties uit te voeren en zo meer winst te maken. Toch is dit geen eenvoudige taak. De aandelenmarkt wordt namelijk gekenmerkt door zijn complexiteit omwille van de talloze factoren die de beurskoersen beïnvloeden. Zo zijn er onder andere economische, politieke en psychologische factoren die een rol spelen.

Sinds de jaren zestig wordt onderzocht of de beurskoersen voorspeld kunnen worden. Eugene Fama was één van de eerste onderzoekers die met eigen theorieën kwam over de voorspelbaarheid van de beurs [5]. De *random walk hypothesis* stelt dat veranderingen in de beurskoersen geen geheugen hebben. Dit zou impliceren dat historische data niet gebruikt kan worden om toekomstige beurskoersen te voorspellen. De *efficient market hypothesis* (EMH) stelt dat alle informatie over een aandeel vervat zit in de prijs van dat aandeel. Dit zou eveneens impliceren dat het onmogelijk is om de beurskoersen te gaan voorspellen. EMH wordt opgedeeld in drie vormen: de zwakke, de semi-sterke en de sterke, afhankelijk van wat er verstaan wordt met het begrip 'alle informatie over een aandeel'. Afhankelijk van de vorm van EMH die als waar aanzien wordt, zou het toch mogelijk zijn om de beurs te voorspellen. Fama ontving voor zijn onderzoek de Nobelprijs economie in 2013.

Twee grote strekkingen onderscheiden zich binnen de systemen die de beursvoorspelling ook in praktijk willen brengen. Een eerste groep gebruikt fundamentele analyse. Ze gaan hierbij proberen om de intrinsieke waarde van een bedrijf te bepalen om vandaar te kijken of de aandelen van dat bedrijf te hoog of te laag genoteerd staan. Afhankelijk hiervan wordt er besloten om aandelen te verkopen of te kopen. De tweede strekking gebruikt technische analyse. Zij proberen om vanuit historische beursdata, zoals aandelenkoersen en verhandelde volumes, de beurskoers te gaan voorspellen. Ze proberen hierbij trends in de beursdata te vinden. Bovendien gebruiken ze technische indicatoren die hen helpen om goede investeringen te vinden.

Eind jaren '90 is men beginnen onderzoeken of nieuwsberichten ook een invloed konden spelen op beurskoersen. Zij bevatten namelijk een schat van informatie over bedrijven. Later werden ook andere tekstuele bronnen zoals financiële jaarverslagen en Twitterberichten toegevoegd aan het onderzoek. Het onderzoek in dit domein wordt verder aangemoedigd door de steeds toenemende databronnen die beschikbaar zijn op het World Wide Web.

Dit artikel begint in sectie 2 met een bespreking van eerder ontwikkelde voorspellingssystemen voor beurskoersen. Sectie 3 gaat dieper in op wat sentimentanalyse kan betekenen voor de voorspelling van beurskoersen. De rest van het artikel probeert te achterhalen welke invloed Twitter op de beurskoersen uitoefent.

Sectie 4 beschrijft het systeem dat ontworpen werd, waarna sectie 5 de experimenten uitlegt die met dat systeem gedaan werden, evenals de resultaten van deze experimenten. Het artikel besluit met een conclusie in sectie 6.

2. Bestaande voorspellingssystemen

De eerste systemen die een poging deden om de beurs te voorspellen werden ontworpen aan de Hong Kong University of Science and Technology. Cho [3], Leung [10] and Peramunetilleke [13] trachtten elk in hun werk om met behulp van nieuwsberichten en probabilistische regels de Hang Seng Index te voorspellen.

Vele andere systemen volgden hierna. De makers van deze systemen vinden echter geen consensus over welke methode de beste oplossing levert. Zo gebruiken Lavrenko [8] en Gidófalvi [7] een *naive Bayes classifier*, terwijl Fung [6] en Schumacher et al. [14–20] *support vector machines* (SVMs) gebruiken in hun werk. Gidófalvi vond in zijn werk dat er een *window of influence* van 20 minuten bestond, dit betekent dat een nieuwsbericht een invloed heeft op de beurskoers van een bedrijf 20 minuten voor het gepubliceerd wordt tot 20 minuten erna.

Later werden er naast nieuwsberichten ook andere tekstuele bronnen gebruikt. Zo beschrijft Lee [9] hoe hij met behulp van het hiërarchisch clusteren van *features* afkomstig van financiële jaarverslagen erin slaagt om beurstrends te voorspellen. Ook Twitter is een populaire databron voor voorspellingssystemen. Hierbij wordt er veel belang gehecht aan het sentiment van de tweets. Pak & Paroubek [12] proberen om met behulp van *n-grams* en *POS-tags* een *naive Bayes classifier* te trainen die het sentiment van tweets kan bepalen. Bollen et al. [1] gebruiken OpinionFinder en Google-Profile of Mood States (GPOMS) om de collectieve gemoedstoestand van de Twittergebruikers te bepalen. Van daaruit proberen ze dan de Dow Jones Industrial Average (DJIA) te voorspellen. Anderen proberen meer inzichten te krijgen in het woordgebruik in tweets om zo de beurs te voorspellen. Chen & Lazer [2] gebruiken hiervoor SentiWordNet[1], Zhang et al. [22] tellen emotionele woorden en Evangelopoulos et al. [4] gebruiken *latent semantic analysis*. Smailović et al. [21] trainen een SVM met *bigrams* om zo de beurskoers van bedrijven te voorspellen.

Hoewel de meeste onderzoekers beweren dat ze goede resultaten behalen en dat ze winsten zouden kunnen maken met hun systeem, is dit laatste hoogst twijfelachtig. Onderzoekers houden namelijk vaak geen rekening met transactiekosten, terwijl deze zeer hoog kunnen oplopen.

3. Sentimentanalyse

Sentimentanalyse is de taak die probeert om opinies, emoties en polariteit te detecteren, te extraheren of samen te vatten. Het steunt hierbij op *features* die door hun aan- of afwezigheid het sentiment van een tekst kunnen aangeven [11]. De taak bestaat uit drie delen. Het eerste deel focust zich op het verzamelen van een goede dataset. Soms kan het nodig zijn dat deze data (deels) gelabeld is. Vervolgens moet er bepaald worden met welke *features* er gewerkt zal worden. Er wordt hierbij onderscheid gemaakt tussen vier categorieën: syntactische, semantische, *link-based* en stilistische *features*. Tot slot wordt de eigenlijke analyse gedaan door een algoritme. Er zijn verschillende mogelijkheden: *ad hoc rule based approaches*, *supervised learning approaches* en *semi-supervised learning approaches*. Het voordeel aan *semi-supervised* tegenover *supervised* is dat het minder gelabelde data nodig heeft. Het labelen van data is vaak een dure en arbeidsintensieve taak.

Sentimentanalyse kan helpen om meer inzichten te krijgen in de invloed van berichten op de beurs. Een gedetailleerder onderscheid tussen nieuws met een positieve connotatie en nieuws met een negatieve, laat ook toe om de verbanden met de beurs in meer detail te bestuderen. Bovendien geeft het de kans om sneller uit te maken welke berichten nuttig zijn en welke minder.

4. Beschrijving van het systeem

Het systeem verzamelt eerst relevante documenten. Hier zijn dat tweets die informatie bevatten over bedrijven of over producten van die bedrijven. Er werden twee datasets gebruikt: één met manueel gelabelde tweets, afkomstig van Sanders Analytics[2] en één met finaniciële tweets, afkomstig van Infochimps[3]. Financiële tweets kunnen herkend worden aan het gebruik van een dollar symbool, gevolgd door de ticker van een bedrijf.

Alvorens de tweets bruikbaar zijn voor experimenten moeten ze eerst gepreprocessed worden. Tijdens de *preprocessing* worden eerst woorden verwijderd die niet bijdragen tot het sentiment van een tweet, zoals bijvoorbeeld tijdstippen of gebruikersnamen. Vervolgens worden typische curiositeiten van Twitter aangepast. Zo worden woorden die als *hashtag* bij een tweet

[1] http://sentiwordnet.isti.cnr.it/

[2] www.sananalytics.com

[3] www.infochimps.com

staan vervangen door hetzelfde woord zonder het #-symbool. Tot slot wordt ervoor gezorgd dat uniformiteit bekomen wordt onder de verschillende tweets. Zo wordt bijvoorbeeld alles in kleine letters gezet en worden emoticons vervangen door een generisch equivalent. Alle smileys worden bijvoorbeeld vervangen door ”happy_emoticon”.

Vervolgens zal het systeem *features* kiezen die zullen toelaten om het sentiment te bepalen. Er wordt in dit systeem gewerkt met een *bag-of-words* representatie van de tweets. Dit is een populaire manier omdat hij zo eenvoudig is. Nadeel is wel dat het geen rekening houdt met grammaticale samenhang of woordvolgorde. Toch wezen eerdere experimenten uit dat het van nut kan zijn voor de analyse van nieuwsartikels.

De laatste stap is het effectief uitvoeren van een algoritme om het sentiment te bepalen. In dit systeem worden er verschillende methodes voor gebruikt, afhankelijk van de experimenten die uitgevoerd moeten worden. Zo wordt er bijvoorbeeld gebruik gemaakt van *k-means*, een algoritme dat gelijkaardige tweets clustert. Nieuwe tweets krijgen dan een sentiment toegewezen afhankelijk van de cluster waartoe ze behoren. De volgende sectie gaat dieper in op de experimenten en de methodes die daarbij gebruikt werden.

5. Experimenten

5.1. Clustering

De eerste reeks experimenten maakten gebruik van clustering om het sentiment te bepalen. Hierbij werd er gebruik gemaakt van de gelabelde data van Apple en Microsoft uit de data set van Sanders Analytics. De data van Google was niet bruikbaar omdat deze tweets bevat uit het weekend, en dan is de beurs gesloten. De data van Twitter is ook onbruikbaar want Twitter was op dat moment nog niet beursgenoteerd. Tabel 1 toont de samenstelling van deze data set. Er kan meteen opgemerkt worden dat er veel meer neutrale tweets zijn dan positieve of negatieve. Dit kwam ook al naar voor in andere onderzoeken uit de literatuur.

De data werd eerst geclusterd volgens de manueel toegekende tags. Hierbij werden de tweets gegroepeerd in intervallen van 1, 5, 10 en 20 minuten. Bovendien werd er getest wat de reactietijd was tussen de tweets en de veranderingen van de beurs. Het verschijnen van positieve, negatieve en neutrale tweets geeft aanleiding om een sentimentcurve op te stellen, die dan gecorreleerd kan worden aan de beurskoersen.

Tabellen 2 en 3 tonen dat de resultaten van dit experiment weinig significant waren. Bovendien kwamen de verschillende experimenten niet tot een veralgemeenbare correlatie. Toch zijn er op de verschillende grafieken tijdelijke correlaties tussen de sentimentcurve en de beurskoersen op te merken. Figuur 1 toont bijvoorbeeld de grafieken van het experiment met de gelabelde data van Apple, voor tweets die gegroepeerd werden per 10 minuten en een reactietijd van de beurs van 10 minuten. Rond het middaguur op 18 oktober (de tweede dag) is er een sterke gelijklopende trend tussen het sentiment uit de tweets en de beurskoers van Apple.

Tabel 1. Samenstelling van de tweets in de dataset van Sanders Analytics

Bedrijf	Aantal tweets			Totaal
	Positief	Negatief	Neutraal	
Apple	150	301	495	946
Google	193	49	560	802
Microsoft	89	124	609	822
Twitter	52	60	532	644
Totaal	484	534	2196	3114

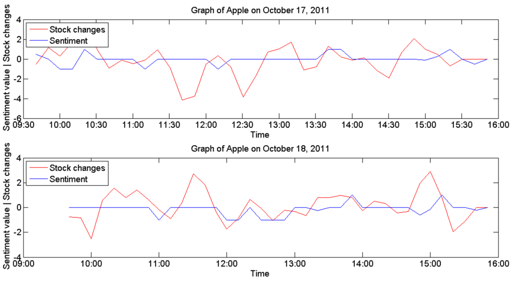
Tabel 2. Correlaties tussen de gelabelde tweets en de beurskoers van Apple

Tweets interval	Reactietijd beurs			
	20 min	10 min	5 min	1 min
20 min	0.3461*	-	-	-
10 min	0.1184	0.1957	-	-
5 min	-0.0420	0.0412	0.0844	-
1 min	-0.0516	-0.0463	-0.0436	-0.0034
$*$: $p-waarde<0.05$				

Tabel 3. Correlaties tussen de gelabelde tweets en de beurskoers van Microsoft

Tweets interval	Reactietijd beurs			
	20 min	10 min	5 min	1 min
20 min	0.0547	-	-	-
10 min	0.0355	-0.1589	-	-
5 min	0.0601	0.0437	-0.1265	-
1 min	0.0582	0.0485	0.0155	-0.0959
$*$: $p-waarde<0.05$				

Hierna werden er nog meer experimenten gedaan die gebruik maakten van clustering. Ditmaal werd er



Figuur 1. Grafiek van Apple op 17-18 oktober 2011

geclusterd met behulp van het *k-means* algoritme, dat gelijkaardige tweets clustert volgens het woordgebruik in de tweets. Opnieuw varieerde het interval waarbinnen tweets gegroepeerd werden en de reactietijd van de beurs tijdens de verschillende experimenten. Ook het aantal clusters was variabel; de experimenten werden uitgevoerd voor 3, 5, 7 en 9 clusters.

De resultaten tonen ditmaal meer significante resultaten. Het resultaat voor 3 clusters kan teruggevonden worden in tabellen 4 en 5. Vergelijkbare resultaten komen terug voor experimenten met meer dan 3 clusters. De tweets van Microsoft zijn bijna altijd positief gecorreleerd aan de beurskoers, terwijl die van Apple een negatieve correlatie tonen.

De experimenten scoren bijna altijd een significant resultaat wanneer de tweets gegroepeerd worden per 10 minuten en de reactietijd van de beurs 10 minuten bedraagt. Tabel 6 bevestigt dit: 5 van de 8 experimenten waren significant met een p-waarde kleiner dan 0.05, 2 anderen hadden een p-waarde kleiner dan 0.1.

Tabel 4. Correlatie tussen geclusterde tweets en de beurskoers van Microsoft voor 3 clusters

Tweets interval	Reactietijd beurs			
	20 min	10 min	5 min	1 min
20 min	0.3836	-	-	-
10 min	0.3470*	0.3559*	-	-
5 min	0.0648	0.1408	0.0415	-
1 min	0.0453	0.0702	0.0736	-0.0944
$*$: $p{-}waarde{<}0.05$				

5.2. Topic modellering

Het volgende experiment had als doel te kijken waarom de beurs soms positief gecorreleerd was en soms negatief. Er werd geprobeerd om te achterhalen welke onderwerpen positieve, negatieve en neutrale tweets kenmerken. Hiervoor werd er *latent Dirichlet allocation* (LDA) gebruikt. Dit staat toe om aan de hand van een *bag-of-words* representatie topics te bepalen die een set van documenten kenmerken. Er werd een LDA met 20 *topics* uitgevoerd op de positieve, negatieve en neutrale tweets van de 4 bedrijven van de dataset van Sanders Analytics en één voor alle positieve, negatieve en neutrale tweets samen.

Tabel 7 toont de eerste 5 *topics* van elke categorie voor de experimenten die alle tweets samennemen. We zien dat de meeste *topics* meningen bevatten van gebruikers over bedrijven en producten. Zo zien we bijvoorbeeld verschillende mobiele besturingssystemen opduiken in de *topics*: iOS (Apple), Mango (Microsoft) en Android (Google). Dit ligt in de lijn van de verwachtingen: Twitter is een populair medium voor gebruikers om hun meningen te uiten en ervaringen te delen. Sommige woorden komen zowel bij positieve, negatieve als neutrale tweets voor. Zo komt het woord ’update’ voor bij elke groep van Apple. Dit komt door de verschillende ervaringen van gebruikers bij bepaalde producten of services. Verschillende woorden binnen de *topics* verraden of deze ervaringen positief of nega-

Tabel 5. Correlatie tussen geclusterde tweets en de beurskoers van Apple voor 3 clusters

Tweets interval	Reactietijd beurs			
	20 min	10 min	5 min	1 min
20 min	-0.2418	-	-	-
10 min	-0.2408*	-0.2548*	-	-
5 min	-0.1253	-0.1507	-0.1137	-
1 min	-0.0937*	-0.1200*	-0.0965*	-0.0261
$*$: $p{-}waarde{<}0.05$				

Tabel 6. Samenvatting van de correlaties voor tweets gegroepeerd per 10 minuten en een reactietijd van de beurs van 10 minuten

Aantal clusters	Apple	Microsoft
3	-0.2548*	0.3559*
5	-0.2392*	0.2862^1
7	-0.2318*	0.2380
9	-0.1917^2	0.3247*
$*{:}p{-}waarde{<}0.05$ $^1{:}p{-}waarde{=}0.0859$ $^2{:}p{-}waarde{=}0.0972$		

tief waren. Positieve ervaringen worden bijvoorbeeld omschreven met ’loving’ en ’appreciating’, terwijl negatieve ervaringen gecombineerd worden met woorden als ’fix’ en ’annoyed’. Dit soort woorden zou met behulp van een thesaurus automatisch gedetecteerd kunnen worden, een aanpak die Chen en Lazer vroeger ook al toegepast hebben met SentiWordNet.

Tabel 7. Eerste 5 topics van de LDA op alle tweets

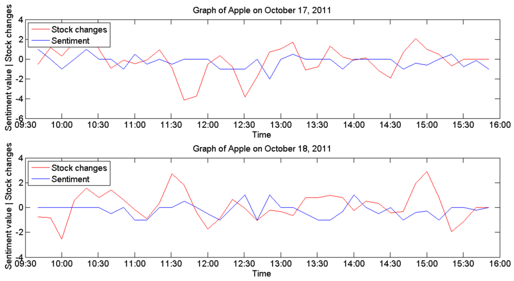
	Topic nummer	Meest belangrijke woorden
positief	1	ios, better, mango, right, upgrade, down, guy, info, live
	2	siri, it, technology, actually, add, thx, happy, support, things
	3	shows, cant, look, make, new, feel, little, blown, thats
	4	android, google, ics, sandwich, ice, galaxynexus, data, looking, doing, idea
	5	rt, users, people, service, facebook, ive, wait, way, didnt
negatief	6	microsoft, apple, fail, windows, hate, day, steve, eclipsed, skype, people
	7	issues, need, thanks, make, hold, software, buy, hell, youre
	8	ios, problem, not, accounts, report, thats, think, wont, annoyed
	9	store, service, customer, genius, apps, life, sucks, w, using
	10	new, fucking, work, app, rt, ipad, like, hey, fuck
neutraal	11	little, newly, xbla, coming, technological, announces, onl, kinda, learned
	12	twits, im, knololi, followers, neea, yoo, wannabe, leopard, autopilot
	13	google, searay, got, mobil, digital, multitouch, bookcase, funny, possibly
	14	ipads, apple, sinking, dont, big, watson, hey, dear, used
	15	going, lovatics, talent, consider, iphoneartists, let, more, beat, bus

5.3. Stanford sentiment treebank

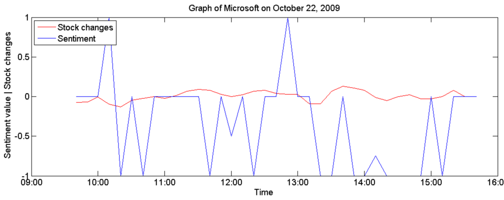
Een laatste groep van experimenten werd uitgevoerd met de Stanford sentiment treebank. Dit kijkt naar de structuur van zinnen en bepaalt dan het sentiment van deze zinnen. Het grote voordeel is dat er geen gelabelde data meer nodig is. De experimenten werden uitgevoerd door tweets te groeperen in intervals van 10 minuten en een reactietijd van 10 minuten te nemen, aangezien de eerdere experimenten uitwezen dat dit de beste resultaten gaf.

Het experiment werd eerst uitgevoerd op dezelfde data als eerdere experimenten. De resultaten zijn niet significant, hoewel ze weer een positieve correlatie voor Microsoft en een negatieve correlatie voor Apple teruggeven. De grafieken tonen ook weer correlaties op bepaalde ogenblikken van de dag. De grafiek van Apple staat weergegeven in figuur 2 en toont dezelfde correlaties als figuur 1 tijdens de middaguren van 18 oktober.

Het experiment werd ook gedaan voor de tweets van Microsoft op de eerste dag dat Windows 7 in de winkels lag. Deze dag was namelijk de dag met de meeste tweets uit de hele dataset. Ook hier werd geen significant resultaat gevonden. Toch waren er opmerkelijke resultaten gedurende de dag zelf. Zo waren de eerste 10 intervallen van 10 minuten gecorreleerd met een waarde van 0.6420, terwijl er vanaf 11 uur eveneens 10 intervallen significant gecorreleerd waren met een waarde van -0.5247. De grafiek van dit experiment staat in figuur 3.



Figuur 2. Grafiek van Apple op 17-18 oktober 2011



Figuur 3. Grafiek van Microsoft op 22 oktober 2009

6. Conclusie

Dit artikel onderzocht of Twitter één van de factoren is die een invloed heeft op de beurs. Uit de verschillende experimenten is gebleken dat de invloed van Twitter varieert in de tijd: soms zijn er sterke verbanden tussen tweets en de beurskoersen, soms zijn er geen verbanden. Voor beleggers is het natuurlijk belangrijk om deze momenten zo snel mogelijk op te merken.

Uit de resultaten van de LDA is gebleken dat individuele woorden het sentiment van een tweet kunnen verraden, daarom is een bag-of-words representatie een goede manier om tweets voor te stellen in voorspellingssystemen.

Vanuit de data die gebruikt werd voor dit artikel, zou er geconcludeerd kunnen worden dat er een *window of influence* van 10 minuten zou kunnen bestaan tussen de verschijning van tweets en hun effect op de beurskoersen.

Referenties

[1] Bollen, J.; Mao, H.; Zeng, X. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[2] Chen, R.; Lazer, M. Sentiment Analysis of Twitter Feeds for the Prediction of Stock Market Movement. 2011.

[3] Cho, V.; Wüthrich, B.; Zhang, J. Text Processing for Classification, 1998.

[4] Evangelopoulos, N.; Magro, M. J.; Sidorova, A. The Dual Micro/Macro Informing Role of Social Network Sites: Can Twitter Macro Messages Help Predict Stock Prices? *Informing Science*, 15:247–268, 2012.

[5] Fama, E. F. Random Walks in Stock Market Prices. *Financial Analysts Journal*, 21(5):55–59, September/October 1965.

[6] Fung, G. P. C.; Yu, J. X.; Lam, W. News Sensitive Stock Trend Prediction. In *Advances in Knowledge Discovery and Data Mining*, volume 2336 of *Lecture Notes in Computer Science*, pages 481–493. Springer Berlin Heidelberg, 2002.

[7] Gidófalvi, G. Using news articles to predict stock price movements, June 2001.

[8] Lavrenko, V.; Schmill, M.; Lawrie, D.; Ogilvie, P.; Jensen, D.; Allan, J. Language Models for Financial News Recommendation. In *Proceedings of the Ninth International Conference on Information and Knowledge Management*, pages 389–396, 2000.

[9] Lee, A. J. T.; Lin, M.-C.; Kao, R.-T.; Chen, K.-T. An Effective Clustering Approach to Stock Market Prediction. In *Pacific Asia Conference on Information Systems, PACIS 2010, Taipei, Taiwan, 9-12 July 2010*, pages 345–353, 2010.

[10] Leung, S. K. F. Automatic stock market: predictions from World Wide Web data. Master's thesis, The Hong Kong University of Science and Technology, 1997.

[11] Moens, M.-F.; Li J.; Chua T.-S. *Mining user generated content*. Chapman and Hall/CRC, 2014. ISBN-13: 9781466557406.

[12] Pak, A.; Paroubek, P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 1320–1326, May 2010.

[13] Peramunetilleke, D. C. A system for exchange rate forecasting using news headlines. Master's thesis, The Hong Kong University of Science and Technology, 1997.

[14] Schumaker, R. P. An Analysis of Verbs in Financial News Articles and Their Impact on Stock Price. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, WSA '10, pages 3–4. Association for Computational Linguistics, 2010.

[15] Schumaker, R. P. Analyzing Parts of Speech and their Impact on Stock Price. *Communications of the International Information Management Association*, 10(3):1–10, 2010.

[16] Schumaker, R. P.; Chen, H. Textual Analysis of Stock Market Prediction Using Financial News Articles. In *12th Americas Conference on Information Systems*, August 2006.

[17] Schumaker, R. P.; Chen, H. Evaluating a News-aware Quantitative Trader: The Effect of Momentum and Contrarian Stock Selection Strategies. *Journal of the American Society for Information Science and Technology*, 59(2):247–255, January 2008.

[18] Schumaker, R. P.; Chen, H. A Quantitative Stock Prediction System Based on Financial News. *Journal Information Processing and Management*, 45(5):571–583, September 2009.

[19] Schumaker, R. P.; Chen, H. Sentiment Analysis of Financial News Articles. In *20th Annual Conference of International Information Management Association*, October 2009.

[20] Schumaker, R. P.; Chen, H. Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFin Text System. *ACM Transactions on Information Systems*, 27(2):12:1–12:19, March 2009.

[21] Smailović, J.; Grčar, M.; Lavrač, N.; Žnidaršič, M. Predictive Sentiment Analysis of Tweets: A Stock Market Application. In *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data*, volume 7947 of *Lecture Notes in Computer Science*, pages 77–88. Springer Berlin Heidelberg, 2013.

[22] Zhang, X.; Zhou, Y.; Bailey, J.; Ramamohanarao, K. Sentiment Analysis by Augmenting Expectation Maximisation with Lexical Knowledge. In *Web Information Systems Engineering - WISE 2012*, volume 7651 of *Lecture Notes in Computer Science*, pages 30–43. Springer Berlin Heidelberg, 2012.

Bibliography

[1] A Norm Al. List of English Stop Words. URL: http://norm.al/2009/04/14/list-of-english-stop-words/, last checked on 2014-05-28.

[2] Abbasi, A.; Chen, T.; Salem, A. Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. *ACM Transactions on Information Systems*, 26(3), June 2008.

[3] Agrawal, R.; Rajagopalan, S.; Srikant, R.; Xu Y. Mining newsgroups using networks arising from social behavior. In *WWW '03 Proceedings of the 12th international conference on World Wide Web*, pages 529–535, 2003.

[4] Blei, D. M.; Ng, A. Y.; Jordan, M. I. Latent Dirichlet Allocation. *The Journal of Machine Learning Research* , 3:993–1022, March 2003.

[5] Bollen, J.; Mao, H.; Zeng, X. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[6] Brandwatch. Sentiment Analysis. URL: http://www.brandwatch.com/wp-content/uploads/2012/11/Sentiment-Analysis.pdf, last checked on 2014-05-14.

[7] Buche, A.; Chandak, M. B.; Zadgaonkar, A. An Approach for Online Analysis using Expectation Maximization. *International Journal of Innovative Research in Computer and Communication Engineering*, 1(4):1064–1071, June 2013.

[8] Chapelle, O.; Schölkopf, B.; Zien, A. *Semi-Supervised Learning*. The MIT Press, 2006. ISBN-13: 9780262033589.

[9] Chen, R.; Lazer, M. Sentiment Analysis of Twitter Feeds for the Prediction of Stock Market Movement. 2011.

[10] P. Chesleyn. Using verbs and adjectives to automatically classify blog sentiment. In *In Proceedings of AAAI-CAAW-06, the Spring Symposia on Computational Approaches*, pages 27–29, 2006.

[11] Cho, V.; Wüthrich, B.; Zhang, J. Text Processing for Classification, 1998.

[12] CIA. The World Factbook. URL: https://www.cia.gov/library/publications/the-world-factbook/fields/2195.html, last checked on 2014-05-11.

[13] Cui, H.; Mittal, V.; Datar, M. Comparative Experiments on Sentiment Classification for Online Product Reviews. In *Proceeding AAAI'06 proceedings of the 21st national conference on Artificial intelligence - Volume 2*, pages 1265–1270, 2006.

[14] Efron, M. Cultural Orientation: Classifying Subjective Documents by Cociation Analysis. In *Proceedings of the AAAI Fall Symposium on Style and Meaning in Language, Art, Music, and Design*, pages 41–48, 2004.

[15] Ekman, P. An argument for basic emotions. *Cognition and emotion*, 6(3/4):169–200, 1992.

[16] Evangelopoulos, N.; Magro, M. J.; Sidorova, A. The Dual Micro/Macro Informing Role of Social Network Sites: Can Twitter Macro Messages Help Predict Stock Prices? *Informing Science*, 15:247–268, 2012.

[17] Fama, E. F. Random Walks in Stock Market Prices. *Financial Analysts Journal*, 21(5):55–59, September/October 1965.

[18] Fei, Z.;Liu, J.; Wu, G. Sentiment classification using phrase patterns. In *Computer and Information Technology, 2004. CIT '04. The Fourth International Conference on*, pages 1147–1152, 2004.

[19] Fung, G. P. C.; Yu, J. X.; Lam, W. News Sensitive Stock Trend Prediction. In *Advances in Knowledge Discovery and Data Mining*, volume 2336 of *Lecture Notes in Computer Science*, pages 481–493. Springer Berlin Heidelberg, 2002.

[20] Gamallo, P.; Garcia, M.; Fernández-Lanza, S. TASS: A Naive-Bayes strategy for sentiment analysis on Spanish tweets. In *XXIX Congreso de la Sociedad Española de Procesamiento de lenguaje natural. Workshop on Sentiment Analysis at SEPLN*, pages 126–132, 2013.

[21] Gamon, M. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In *Proceeding of COLING-04, the 20th International Conference on Computational Linguistics*, pages 841–847, 2004.

[22] Gidófalvi, G. Using news articles to predict stock price movements, June 2001.

[23] Griffiths, T.; Steyvers, M. Finding Scientific Topics. In *Proceedings of the National Academy of Sciences*, volume 101.

[24] He, Y.; Zhou, D. Self-training from labeled features for sentiment analysis. *Information Processing & Management*, 47(4):606–616, 2011.

[25] Hong, L.; Davison, B. D. Empirical Study of Topic Modeling in Twitter. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 80–88. ACM, 2010.

[26] Investopedia. Complete Guide To Corporate Finance. URL: http://www.investopedia.com/walkthrough/corporate-finance/, last checked on 2014-05-07.

[27] Investopedia. Fundamental Analysis: Introduction. URL: http://www.investopedia.com/university/fundamentalanalysis/, last checked on 2014-05-11.

[28] Investopedia. Technical Analysis: Introduction. URL: http://www.investopedia.com/university/technical/, last checked on 2014-05-11.

[29] Lavrenko, V.; Schmill, M.; Lawrie, D.; Ogilvie, P.; Jensen, D.; Allan, J. Language Models for Financial News Recommendation. In *Proceedings of the Ninth International Conference on Information and Knowledge Management*, pages 389–396, 2000.

[30] Lee, A. J. T.; Lin, M.-C.; Kao, R.-T.; Chen, K.-T. An Effective Clustering Approach to Stock Market Prediction. In *Pacific Asia Conference on Information Systems, PACIS 2010, Taipei, Taiwan, 9-12 July 2010*, pages 345–353, 2010.

[31] Leung, S. K. F. Automatic stock market: predictions from World Wide Web data. Master's thesis, The Hong Kong University of Science and Technology, 1997.

[32] MacQueen, J. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297. University of California Press, 1967.

[33] MathWorks. K-means Clustering. URL: http://www.mathworks.nl/help/stats/k-means-clustering.html, last checked on 2014-06-01.

[34] MathWorks. Linear or rank correlation. URL: http://www.mathworks.nl/help/stats/corr.html, last checked on 2014-06-01.

[35] Mishkin, F. S. . *The economics of money, banking, and financial markets.* Addison-Wesley, 2004. ISBN-13: 9780345391803.

[36] Moens, M.-F.; Li J.; Chua T.-S. *Mining user generated content.* Chapman and Hall/CRC, 2014. ISBN-13: 9781466557406.

[37] Mullen, T.; Collier, N. Sentiment Analysis using Support Vector Machines with Diverse Information Sources. In *Proceedings of EMNLP-04, 9th Conference on Empirical Methods in Natural Language Processing*, volume 4, pages 412–418, 2004.

[38] Nebraska Council. Top 2000 Valued Companies with Ticker Symbols. URL: http://www.nebraskacouncil.org/smg/documents/Top%202000%20Valued%20Companies%20with%20Ticker%20Symbols.pdf, last checked on 2014-05-17.

[39] Pak, A.; Paroubek, P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 1320–1326, May 2010.

[40] Pang, B.; Lee, L.; Vaithyanathan, S. Thumbs Up?: Sentiment Classification Using Machine Learning Techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, pages 79–86, 2002.

[41] Peramunetilleke, D. C. A system for exchange rate forecasting using news headlines. Master's thesis, The Hong Kong University of Science and Technology, 1997.

[42] Perry, M. J. World stock market capitalization closes year at $54.6 trillion. URL: http://www.aei-ideas.org/2013/01/world-stock-market-capitalization-at-54-6-trillion/, last checked on 2014-06-02.

[43] Prasad, S. Micro-blogging Sentiment Analysis Using Bayesian Classification Methods. 2010.

[44] Riloff, E.; Patwardhan, S.; Wiebe, J. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the Seventh Conference on Natural Language Learning (CoNLL-2003)*, pages 25–32, 2003.

[45] Saif, H.; He Y.; Alani, H. Semantic sentiment analysis of Twitter . In *Proceeding ISWC'12 Proceedings of the 11th international conference on The Semantic Web*, pages 508–524, 2012.

[46] Schumaker, R. P. An Analysis of Verbs in Financial News Articles and Their Impact on Stock Price. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, WSA '10, pages 3–4. Association for Computational Linguistics, 2010.

[47] Schumaker, R. P. Analyzing Parts of Speech and their Impact on Stock Price. *Communications of the International Information Management Association*, 10(3):1–10, 2010.

[48] Schumaker, R. P.; Chen, H. Textual Analysis of Stock Market Prediction Using Financial News Articles. In *12th Americas Conference on Information Systems*, August 2006.

[49] Schumaker, R. P.; Chen, H. Evaluating a News-aware Quantitative Trader: The Effect of Momentum and Contrarian Stock Selection Strategies. *Journal of the American Society for Information Science and Technology*, 59(2):247–255, January 2008.

[50] Schumaker, R. P.; Chen, H. A Quantitative Stock Prediction System Based on Financial News. *Journal Information Processing and Management*, 45(5):571–583, September 2009.

[51] Schumaker, R. P.; Chen, H. Sentiment Analysis of Financial News Articles. In *20th Annual Conference of International Information Management Association*, October 2009.

[52] Schumaker, R. P.; Chen, H. Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFin Text System. *ACM Transactions on Information Systems*, 27(2):12:1–12:19, March 2009.

[53] Setiady, J.; Maharani, W.; Rismala, R. Feature-Based Sentiment Analysis in Online Review with Semi-Supervised Support Vector Machines. In *Proceedings of the Information Systems International Conference (ISICO)*, pages 314–319, 2013.

[54] Smailović, J.; Grčar, M.; Lavrač, N.; Žnidaršič, M. Predictive Sentiment Analysis of Tweets: A Stock Market Application. In *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data*, volume 7947 of *Lecture Notes in Computer Science*, pages 77–88. Springer Berlin Heidelberg, 2013.

[55] Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C.; Ng, A.; Potts, C. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*. 2013.

[56] Twitter. About. URL: https://about.twitter.com/company, last checked on 2014-06-02.

[57] Twitter. Developer Rules of the Road. URL: https://dev.twitter.com/terms/api-terms, last checked on 2014-05-19.

[58] Twitter. REST API Rate Limiting in v1.1. URL: https://dev.twitter.com/docs/rate-limiting/1.1, last checked on 2014-05-19.

[59] Twitter. Tweets. URL: https://dev.twitter.com/docs/platform-objects/tweets, last checked on 2014-05-17.

[60] Twitter. Users. URL: https://dev.twitter.com/docs/platform-objects/users, last checked on 2014-05-18.

[61] Twitter. Using the Twitter Search API. URL: https://dev.twitter.com/docs/using-search, last checked on 2014-05-19.

[62] Whitelaw, C.; Garg, N. ; Argamon, S. Using Appraisal Groups for Sentiment Analysis. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pages 625–631, 2005.

[63] Wiebe, J.; Wilson, T.; Bruce, R.; Bell, M.; Martin, M. Learning subjective language. *Computer Linguistics*, 30(3):277–308, 2004.

[64] Wikipedia. Financial market. URL: http://en.wikipedia.org/wiki/Financial_market, last checked on 2014-05-07.

[65] Wikipedia. List of emoticons. URL: http://en.wikipedia.org/wiki/List_of_emoticons, last checked on 2014-05-28.

[66] Wikipedia. Ticker tape. URL: http://en.wikipedia.org/wiki/Ticker_tape, last checked on 2014-05-08.

[67] Wilson, T.; Wiebe, J.; Hoffman, P. Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT '05 Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354, 2005.

[68] World Federation of Exchanges. 2013 WFE Market Highlights. URL: http://www.world-exchanges.org/files/2013_WFE_Market_Highlights.pdf, last checked on 2014-05-09.

[69] Yang, Y; Xu, C.; Ren, G. Sentiment Analysis of Text Using SVM. In *Electrical, Information Engineering and Mechatronics 2011*, volume 138 of *Lecture Notes in Electrical Engineering*, pages 1133–1139. Springer London, 2012.

[70] Yates, A.; Goharian, N.; Yee, W. G. Semi-supervised probabilistic sentiment analysis: Merging labeled sentences with unlabeled reviews to identify sentiment. *Proceedings of the American Society for Information Science and Technology*, 50(1):1–10, 2013.

[71] Zagibalov, T.; Carroll J. Automatic seed word selection for unsupervised sentiment classification of Chinese text. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*, pages 1073–1080, 2008.

[72] Zhang, X; Fuehres, H.; Gloor, P. A. Predicting Stock Market Indicators Through Twitter "I hope it is not as bad as I fear". *Procedia - Social and Behavioral Sciences*, 26(0):55–62, 2011.

[73] Zhang, X.; Zhou, Y.; Bailey, J.; Ramamohanarao, K. Sentiment Analysis by Augmenting Expectation Maximisation with Lexical Knowledge. In *Web Information Systems Engineering - WISE 2012*, volume 7651 of *Lecture Notes in Computer Science*, pages 30–43. Springer Berlin Heidelberg, 2012.

[74] Zhou, G.; Joshi, H.; Bayrak, C. Topic Categorization for Relevancy and Opinion Detection. In *Proceedings of the 16th Text Retrieval Conference*, 2007.

Fiche masterproef

Student: Jeroen Ruytings

Titel: Predicting the stock market by using sentiment analysis on Twitter

Nederlandse titel: Het voorspellen van de beurs met behulp van sentimentanalyse op Twitter

UDC: 681.3

Korte inhoud:

Stock market investors would have a great help if a system would tell them if the stock market will go up or down in the future. It would give them the opportunity to invest smartly and gain much more profit. However very desired, such a system is not easy to make. The stock market is characterized by its complexity, with economic, political and psychological factors influencing the stock movements. This master thesis tries to find out if Twitter is also part of those factors that have an influence on the stock market. The thesis starts with a literature study on previous work. Multiple researchers have studied the problem, using news articles, financial reports or microblogging posts to predict the stock market. The researchers do not agree on the best technique to predict the stock market, but they all claim to have promising results, showing correlations between their used source and the stock market fluctuations. The reality shows that their systems are rarely making profit, maybe due to the fact that the stock market is very complex and does not depend on a single factor. Different experiments were done to get a better understanding of Twitter's role in stock market prediction. The experiments show that a window of influence of 10 minutes could exist, meaning a period for which tweets might have an impact on stock prices. From the experiments with the data used in this thesis, no total returning correlation between tweets and stock market in a whole day was found. Nevertheless, for some moments of the day the correlation between tweets and stock market becomes significant for a longer period of time. This might suggest that the influence of Twitter on the stock market varies during the different time periods (days, weeks or even years).

Thesis voorgedragen tot het behalen van de graad van Master of Science in de ingenieurswetenschappen: computerwetenschappen, hoofdspecialisatie Artificiële intelligentie

Promotoren: Prof. dr. M.-F. Moens
Prof. dr. ir. M. Van Barel

Assessoren: Dr. ir. J. Ramon
Dr. O. Kolomiyets

Begeleider: Dr. J. C. Gomez