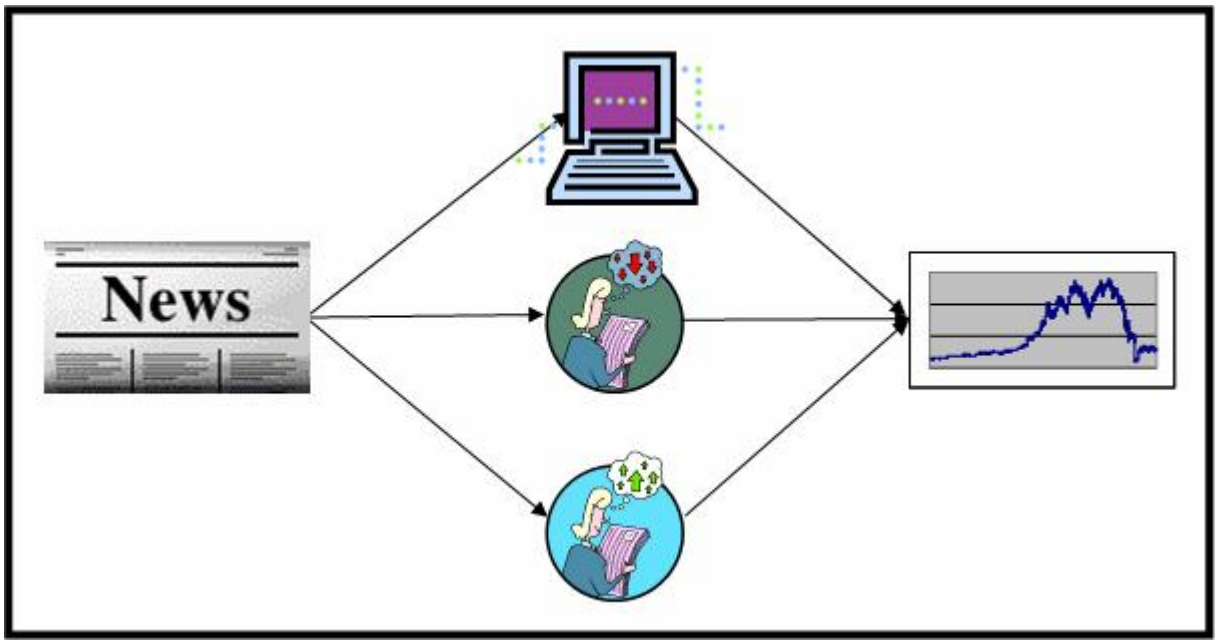
Master Thesis

Interpreting News Flashes for Automatic Stock Price Movement Prediction



Written by:
Fook Hwa Tan

Supervised by:
dr. ir. Jan van den Berg

Informatics & Economics
Faculty of Economics
Erasmus University Rotterdam

Abstract

In this thesis a study is done on the effect or influence of news on the market. It is opted that being able to interpret news messages could assist in predicting market movement in general and stock price movement or trading volume behaviour in particular. It was chosen to interpret news by using Collexis® Fingerprinting technology, hoping to capture the core of a news message. A possible solution to find the effect of news was to reduce the problem to a simpler classification problem. Classification was then done by using Support Vector Machines for model training. The results attained by testing the model on a separate test set showed to be promising.

Acknowledgments

I would like to express my gratitude to my supervisor dr. ir. Jan van den Berg at the Erasmus University of Rotterdam. His continuous enthusiastic support and guidance, while also providing constructive criticism, has helped me a lot in conducting this research and writing this thesis. His patience and wisdom truly encouraged me to conduct this research with excellence. He is also the one who initially provided me with this interesting topic at a seminar conducted by him in 2004.

Another person I have to thank is Nees Jan van Eck, a classmate and friend. He helped me to check my thesis and keep me sharp on the methodology used to perform this experiment. He also provided new insights and ideas to expand upon what I was already planning to do, making this work more complete.

I am also grateful for the help that Rob Stevense at the Erasmus University of Rotterdam has given me. Thanks to him I was able to get information about the stock prices and the indices dating from 1996-1997 through DataStream©.

Thanks to Reuters© I was able to acquire a reliable dataset for my research. They provided me with access to Reuters Corpus© Volume I, which is a vast dataset with all the news articles that Reuters© published during the period 20-08-1996 to 19-08-1997.

Thanks should also be given to Collexis® Software Company for giving me the licence and training to use their software in this research project. Their support given by their helpdesk was also appreciated.

Finally, I would also like to thank my parents and friends who endured this whole process with me and encouraged me to persevere, so I could finish this study and thereby finishing my master degree.

Table of Contents

4. Experimental Setup

5. Results

6. Conclusion

1 INTRODUCTION

"There are many methods for predicting the future. For example, you can read horoscopes, tea leaves, tarot cards, or crystal balls. Collectively, these methods are known as 'nutty methods.' Or you can put well-researched facts into sophisticated computer models, more commonly referred to as 'a complete waste of time.' "

--Scott Adams (1957 -), The Dilbert Future

"It is far better to foresee even without certainty than not to foresee at all."

--Henri Poincare in *The Foundations of Science*, page 129.

1.1 Motivation

The future is a topic often pondered upon and written about by many people. From as far back in history as we know, people have been fascinated by the future. Even in our day and age, many books and movies are about the future. For thousands of years, people have used whatever was at hand to divine the future in order to fulfil a basic yearning for guidance and control.

Knucklebones, entrails, sticks, stones, shells and many manmade objects have been used for predicting the future. Not only tangible objects are used, but also intangible objects can be used to make predictions of the future, like dreams, smoke, winds and auras. These practices are often called divinations. Divination[1] is defined as the practice of seeking knowledge of the future of the unknown by supernatural means. In contrast to fortunetelling, people practicing divination have a less fatalistic view of the world, because they believe that the knowledge acquired gives them the opportunity to change or affect the future. Divination, therefore, gives more room to a human's freewill.

[1] Definition of the noun *divination* is taken from the Compact Oxford English Dictionary of Current English.

A Chinese poet and philosopher of the 6th century BC Lao Tzu said, 'Those who have knowledge, don't predict. Those who predict, don't have knowledge'. Does that mean we shouldn't predict? Or does it only mean we need to be prudent in our dealings with predictions? In the business world, predictions and forecasts are often made to support or assist in the decision making process of finding the most profitable or less risky investment or finance projects. An American Humorist Evan Esar said about economists, 'An economist is an expert who will know tomorrow why the things he predicted yesterday didn't happen today.' As predictions may not always be accurate, they do help in what decisions are to be made, when a sense of what is going to happen is given.

To make correct decisions is important in the business world, because one of the foundational assumptions in economic theory is scarcity. And due to this scarcity of resources, being able to allocate your resources in an efficient way to achieve the highest possible returns at the lowest possible risk is necessary, which is one of the main goals economists strive to achieve. In the financial world many theories exist on how an asset should be valued. Prices are assumed to reflect information both historical and future. Historical information could be, for example, the past prices and future information could be the possible future cash flows that an asset is going to generate. This is called the *Efficient Market Hypothesis* (EMH).

Prices in the markets are generally believed to be established through market dynamics. These dynamics are influenced or affected by the behavior of individuals acting in these markets which will determine the value of an asset. An asset is worth as much as someone is willing to pay for it. The willingness of an individual to pay a certain price is, however, not the only influence on the price. An asset also has its own intrinsic value; a minimum price an individual is willing to sell it for.

The price an individual is often willing to pay is dependent on the knowledge or information that individual has on that asset. A person will try to take into account both past, present and possible future information on that said asset. Future information could be the expectations market participants may have or expectations that individual may have.

In this age of information technology, information or actually data is widely accessible. The *World Wide Web* (WWW) or internet is both a blessing and a curse in this regard. Due to the increased availability of data, the availability of useful data or information has increased as well. Many companies publish their quarterly and annual reports on the WWW. Not only the recent financial ratios, but also historical ratios are easily retrievable. Any investor could look them up on the company's website, regardless of time or place.

Various types of information can be found on the internet. Next to the annual reports with various financial ratios of specific companies, due to the computerizing of the major stock exchanges investors can also retrieve all the prices of anything tradable. This means we have a wealth of data on stock and option prices. Many have used these to find patterns in them to get an extra edge while trading in financial markets, technical analysis. Although it is generally assumed that past performance does not guarantee future performance, it can however be a gauge for future performance.

Another type of information which can be found and may affect prices is disclosures or announcements on upcoming events that may affect a specific or group of assets or companies. The Securities Exchange Act of 1934 specifies disclosure requirements under which publicly held companies file reports with the *Securities and Exchange Commission* (SEC). An excerpt from the *New York Stock Exchange Company Manual*[2] outlines this requirement imposed by the Exchanges and the National Association of Securities Dealers:

A corporation whose securities are listed on the New York Stock Exchange Inc. is expected to release quickly to the public any news or information which might reasonably be expected to materially affect the market for those securities. This is one of the most important and fundamental purposes of the listing agreement which each corporation enters into with the Exchange [p. A-18].

[2] Excerpts are not directly taken from the New York Stock Exchange Company Manual, but taken from Patell, J.M., and Wolfson, M.A., 1982, "Good News, Bad News, and the Intraday Timing of Corporate Disclosures," *The Accounting Review*, Vol 57, No. 3 (Jul., 1982), pp. 509-27.

Another excerpt gives an idea of what is meant by the New York Exchange Inc. when it mentions news or information which might reasonably be expected to materially affect the market for those securities:

Annual and quarterly earnings, dividend announcements, acquisitions, mergers, tender offers, stock splits, and major management changes, and any substantive items of unusual or non-recurrent nature are examples of news items that should be handled on an immediate release basis [p. A-22].

Next to these firm-specific announcements which are required to be disclosed by exchanges, other announcements like macroeconomic news have also shown to be important information sources in the determination of the value of a given asset. In an IMF Working Paper by Funke and Matsuda (2002)[3] several empirical studies, like Hardouvelis (1986)[4], Li and Hu (1998)[5] and Sun and Tong (2000)[6] are cited as giving evidence that stocks are sensitive to news either of a financial or macroeconomic nature.

In a paper by Hamburger (2004)[7], the author interviews an ex-day trader and the trader states that he perceives signals from various resources for his daily work, which are Reuters, financial (and political) news from TV (CNBC), the newspapers, colleagues and the internet. As can be seen news is an important resource for a trader. This suggests there is a possible connection between news and the trading behavior of financial traders. At the release of an announcement, traders often respond in a similar manner, either buying or selling in groups. Although there may be a certain degree of herding, whereby traders look at each other's actions to take action, intuitively certain news announcements tend to give an idea of market price movements. What kind of news has what kind of effect on markets has been a subject of study in the past.

[3] Funke, N., and Matsuda, A., 2002, "Macroeconomic News and Stock Returns in the United States and Germany" IMF Working Paper WP/02/239, (IMF Institute).
[4] Hardouvelis, G.A., 1986, "Macroeconomic Information and Stock Prices," First Boston Working Paper Series FB-86-13, (New York: Columbia University).
[5] Li, L, and Hu, Z.F., 1998, "Responses of the Stock Market to Macroeconomic Announcements Across Economic States," IMF Working Paper 98/79 (Washington International Monetary Fund).
[6] Sun, Q., and Tong, W.H.S., 2000, "The Effect of United States Trade Deficit Announcements on the Stock Prices of United States and Japanese Automakers," *Journal of Financial Research*, Vol. 23, No. 1, pp 15-43.
[7] Hamburger, Y., 2004, "The Exceptional Event," Erasmus University Rotterdam, The Netherlands.

What is news? Many people watch the news, but what does someone really consider to be news? News[8] is defined as newly received or noteworthy information about recent events. News has become a global phenomenon. We no longer only hear about what happens in the town or country we live in, we also know about major events all over the world. With the increase of easily accessible news feeds through the internet, we not only get to know about major events all over the world, we also have access to minor events that took place on the other side of the earth.

News is generated faster and faster these days. News reports are published online in real-time, which mean that whatever happens in the world we get to know about it within minutes and sometimes even less after the event. Oftentimes we even have direct coverage with audio and video feeds supporting. Especially during the US elections of 2004 this was made evident. We here in Europe knew exactly how the Americans were voting, which state had finished counting and which hadn't. We also knew where the president was and what he was planning. The speculation or expectations of the press, whether the president was going to give an early victory speech or not, was also made known to the public via the news.

The above is an example of global news, but even local news about shootings or robberies are in abundance. In the newspaper you can even read obituaries, anniversaries and sometimes even birthdays. Some of this news may not have a global impact, but it is important to somebody. Even minor events which occur in abundance in a certain region could reflect or give an idea of the situation in that region. Many robberies for example could indicate populace with financial problems or an area where many rich people reside.

With computers and the internet, we no longer have to wait for printed news in the form of a newspaper. Major newspapers have their own websites and update as fast as the news comes in and sometimes this can even be faster than television news broadcasts. On TV you need to wait for certain times when the news is on, except for CNN of course which has 24-hour news feeds. For certain important news, you'll see news interruptions on TV. Many news services provide online news feeds, which are accessible at all times.

[8] Definition of the noun *news* is taken from the Compact Oxford English Dictionary of Current English.

Having shown that news announcements may have possible effects on traders' behavior or the economy as a whole, it would be reasonable to assume that being able to capture or quantify this relationship will give the ability or at least assist in predicting movements of the market in the future. Prediction in the scientific community, contrary to divination, is normally described as telling about the future within certain error. The error must be small in order to have a meaningful prediction. The ability to be able to determine or predict market movements has been something economists have striven to achieve in the past. Even now many projects are setup by researchers to find patterns in market movements and search for factors that have influence on the market in order to be able to forecast future market movements. In this same line of thought this research was born.

1.2 Research Objective

This research project was born out of the idea that news has effect on the market and economy. This fact is not only intuitively true, but many researchers have confirmed it in their papers. In the next chapter a description of these works will be given.

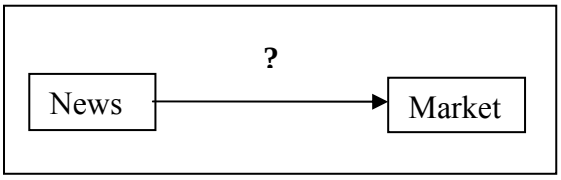


Figure 1: Representation of the conceptual problem.

In this thesis an endeavour is made to find an answer for the following questions: **"What is the effect or influence of news on the market? Does being able to give an interpretation of news messages help in making predictions of market movements?"** A representation of this problem is shown in Figure 1. Although the questions may be simple, the answers are far from simple. The problem is at least two-fold; you have news on one side and the market on the other. Both news and the market are complex entities on their own and are therefore not easily defined, let alone combined.

The question mark above the arrow in Figure 1 could denote both the effect news has on the market as well as the degree of effect news has on the market. This means that it is assumed that news will give the market a push in a certain direction with a certain degree of strength. When a news announcement is released it will cause the market to react or at least it will cause the participants that make up the market to react, because the equilibrium has been disturbed; a shock in the form of new information has been introduced. Different types of news will therefore have different effects on the market. Being able to distinguish the different effects of the different types of news announcements will give the ability to improve market movement predictions. And to have an idea of how the market will move could have far-reaching consequences for risk management, investment decision-making or other economically related decisions.

From another perspective the same problem can be seen in a different light. It is generally assumed that the market drives itself over time. It continuously incorporates new information in its prices trying to achieve equilibrium. News, however, arrive randomly delivering new information to market participants and thereby to the market as a whole on unexpected times. Between the release of an announcement and the market's reaction to it by correcting its prices, a certain amount of time is needed. It will take time for market participants to respond to the new information. This process of the market correcting itself continuously due to external shocks like news is shown in Figure 2. Iturres (2001)[9] shows this evolution of the market in his attempt to find techniques for market prediction.

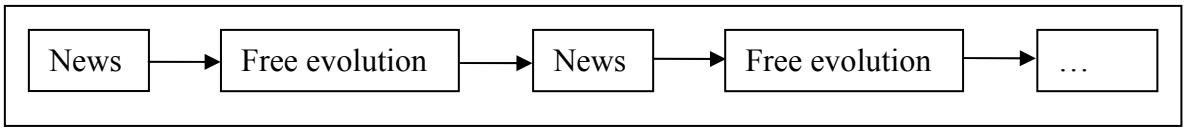


Figure 2: Evolution of the market.

Figure 1 is a representation of the problem from a conceptual level, whereas Figure 2 is a representation of the problem over time; starting at the left and as time progresses moving to the right. In their work, Fleming and Remolona (1999)[10] have shown that U.S. treasury yields respond to news within a short period of time. Especially in this computer age, where

[9] Iturres, A.S., 2001, "Market Prediction Technique," *New Trading Ideas Internet Journal*, Publication 01-07.
[10] Fleming, M.J., and Remolona, E.M., 1999, "Price Formation and Liquidity in the U.S. Treasury Market: The Response to Public Information," *Journal of Finance*, Vol. 54, No. 5, pp. 1901-15.

information is freely and readily available, market participants can respond in a very short time to new information.

1.3 Methodology

To start off this thesis an extensive survey was done on related works. As mentioned earlier the problem is at least two-fold, we have news on one hand and the market on the other. This also means that at least two fields of research must be studied in order to find a suitable solution. One area of research is in economics with a focus in finance, looking at how markets work and studying economic theories. The other area is knowledge discovery or knowledge representation (KR), looking at how news content can be distilled from a message and be made understandable to a computer.

After studying literature in both fields, possible solutions or related works which combine the knowledge of both areas are researched. Descriptions of these existing solutions are given with their assumptions and results. Having shown how others have tried to answer the questions opted in this thesis; a description of the methodology which was used will be given with the reason for this particular choice among the many presented already.

A step by step account of the experimentation process will be given; starting with a description of the datasets used, both for news as well as the financial market. Pre-processing of the datasets will be reported with a step by step account of the experiment.

Finally, the results will be given. A conclusion will then be drawn by an attempt to answer the questions from the previous paragraph. Further suggestions on possible future research will also be given.

1.4 Structure

This research report consists of 6 chapters. In this chapter an introduction has been given. In chapter 2 a report will be given on the research in literature on similar problems. In chapter 3 a description of the methods that are going to be used in this research are given. In chapter 4 an account of the application of the methodology described in chapter 3 is reported. In chapter 5 the results will be shown and in chapter 6 this report will be concluded and some suggestions for future research will be given.

2 LITERATURE

"When you take stuff from one writer it's plagiarism; but when you take it from many writers, it's research."

--Wilson Mizner (1876-1933)

"Those who cannot remember the past are condemned to repeat it."

--George Santayana (1863 - 1952)

2.1 Introduction

A U.S. philosopher George Santayana wrote: "Those who cannot remember the past are condemned to repeat it."[11] In every scientific project a study of related works is essential. We don't want to reinvent the wheel, so to speak. What other researchers have done should be used as a basis for new projects; extending in the areas that they didn't have the time or the resources to study. In some rare cases, they could also lead to brand new ideas. Therefore an extensive survey on available literature on the topic discussed in chapter 1 was done in this thesis.

Before summing up and referring to related works in this chapter on the problem in its entirety, a short account of the background of each area of research as mentioned earlier in the previous chapter will be given; that is the market on one hand and news on the other. The papers incorporating an analysis of both in their research will be discussed thereafter, with their main focus either on the market area or the news area. Finally this chapter will be concluded by drawing a conclusion from the literature surveyed with lessons learnt leading to the methodology that is going to be used in this thesis. In the next chapter then a detailed report will be given on the used methodology.

[11] Santayana, George., US (Spanish-born) philosopher (1863 - 1952), The Life of Reason, Volume 1, 1905.

2.2 Background

In Figure 1 it is suggested that this problem is composed of two concepts, the market and news. News causes the market prices to move and fluctuate, but the market is often also the cause for news reports. This causes therefore a high degree of complexity and that can be seen in the different approaches utilised in the literature surveyed. A start will be made by laying down some background information about the market in general and the EMH in particular. An account of news and its importance to market participants will follow thereafter. Related works focussing on the separate areas will then be given in the section 2.3.

2.2.1 Efficient Market Hypothesis (EMH)[12]

The concept of efficient markets was a chance discovery in 1953 by Maurice Kendall, a British statistician, who presented a controversial paper to the Royal Statistical Society on the behaviour of stock and commodity prices.[13] He suggested that prices of stocks and commodities seem to follow a *random walk*. This means that successive changes in prices are independent. Therefore, no predictable cycles should persist in a series of prices. When however such a cycle would become apparent to investors, they immediately eliminate it by their trading.

You can see why prices in competitive markets must follow a random walk. If past price changes could be used to predict future price changes, investors could make easy profits. In competitive markets prices would adjust immediately through trading until the easy profits from studying past price movements disappeared. As a result, all the information in past prices will be reflected in today's price. The patterns in prices will then cease to exist and price changes in one period will be independent of changes in the next. In other words, prices will follow a random walk.

[12] Brealey, R.A. and Myers, S.C., "*Principles of Corporate Finance*," 7th ed., McGraw-Hill Higher Education, 2003, pp. 344-375.
[13] Kendall, M.G., 1953, "The Analysis of Economic Time Series, Part I. Prices," *Journal of the Royal Statistical Society* 96, pp. 11-25.

Market efficiency is often defined by economists in three levels, distinguished by the information reflected in security prices. In the first level, prices reflect the information contained in the record of past prices; also called the *weak* form of efficiency. The second level of efficiency requires that prices reflect not just past prices but all other published information. This is known as the *semi strong* form of market efficiency. Finally, we have what is known as the *strong* form of efficiency, in which prices reflect all the information from the other two forms and unpublished inside information.

The problem being studied in this thesis will require an analysis of the semi strong form of the efficient-market hypothesis. Researchers have done this by measuring how rapidly security prices respond to different items of news, such as earnings or dividend announcements, news of a takeover, or macroeconomic information. To isolate the effect of an announcement on the price of a stock an event-study[14] can be done on the stock in the months surrounding the announcement. This is, however, a very noisy measure, for the price would also reflect other things that were happening in the market as a whole. A second possibility would be to do a study with a measure of relative performance.

Relative stock return = return on stock – return on market index

This is already better than just looking at returns, but different stocks behave differently to the fluctuations of the market. To adjust for this, past experience might suggest that a change in the market index affected the value of a stock as follows:

Expected stock return = $\alpha + \beta$ x *return on market index*[15]

Alpha (α) states how much on average the stock price changed when the market index was unchanged. Beta (β) tells us how much extra the stock price moved for each 1 percent change in the market index. With this we can calculate the abnormal stock return.

Abnormal stock return = actual stock return – expected stock return

[14] Sar, N.L. van der, 1997, "*Event-studies: Methodologische aspecten,*" Vakgroep Financiering en Belegging, Erasmus Universiteit Rotterdam. [Paper is in Dutch].
[15] This relationship is often referred to as the *market model*.

2.2.2 An Investor's Perspective on News

Having covered in the previous section one of the most basic assumptions in economic theories, the EMH; this section will show the significance of news to market participants in general and a trader in particular.

In a case-study mentioned earlier by Yuri Hamburger[7], an ex-day trader was interviewed. The interviewee mentioned that news announcements were an important driver in the decision making process of a trader in his/her daily trading activity. A general overview of the decision-making process of a trader, from receiving news reports to making the trading decision, is given in Figure 3.

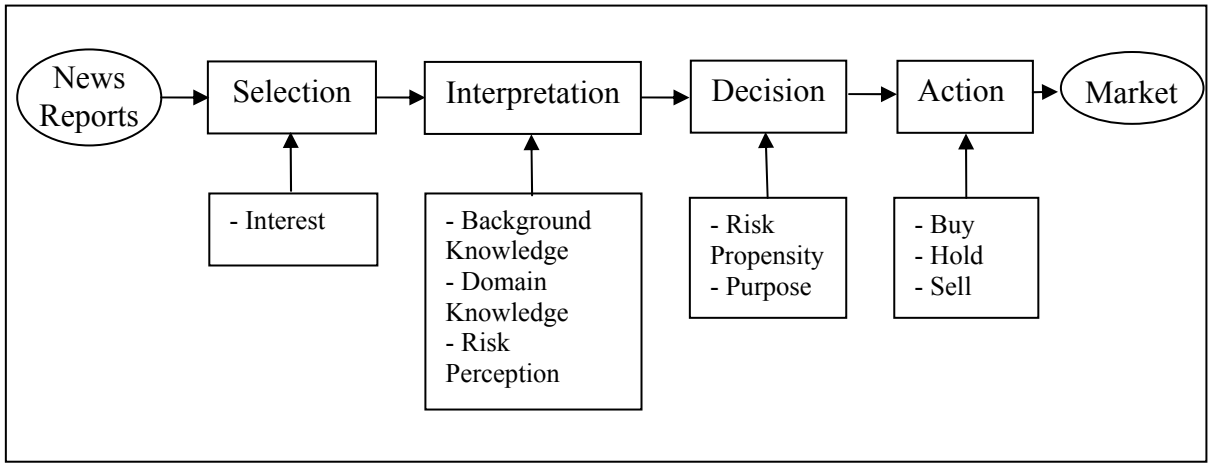


Figure 3: Overview of trader's decision-making process.

News announcements are generated by renowned news services (e.g. Bloomberg, Reuters, CNN, etc.). A trader then makes a selection of interesting news messages. This will either be done due to interest and/or experience. He will make an interpretation of each message depending on his own background, his knowledge on the domain of the topic of the message and his perception of risk. Based on his interpretation he makes a decision, taking into account his propensity to risk and the purpose of his trading. He then offers an order either to buy or sell. It should be noted that being part of the market the action of a single trader may not affect the market depending on the volume traded, but when many traders make the same decision it will have an effect on the market and therefore on the price.

In this day and age, when every process is being automated, one of the problems researchers face in automating this process is finding a representation for news messages which is firstly a satisfactory extraction of the content of the original message and secondly can be understood by computers.

2.2.3 Knowledge Representation (KR)

What is a representation? A representation of content or KR is said by Davis, Shrobe, and Szolovits (1993)[16] to play five important and distinctly different roles, which could explain what it is. They mention that it is not only about the data structure used, but more about the underlying reasons for a chosen method of intelligent reasoning. The five fundamental roles from the paper are as follows:

1. A KR is most fundamentally a surrogate, a substitute for the thing itself, used to enable an entity to determine consequences by thinking rather than acting, i.e., by reasoning about the world rather than taking action in it.
2. It is a set of ontological commitments, i.e., an answer to the question: In what terms should I think about the world?
3. It is a fragmentary theory of intelligent reasoning, expressed in terms of three components: (i) the representation's fundamental conception of intelligent reasoning; (ii) the set of inferences the representation sanctions; and (iii) the set of inferences it recommends.
4. It is a medium for pragmatically efficient computation, i.e., the computational environment in which thinking is accomplished. One contribution to this pragmatic efficiency is supplied by the guidance a representation provides for organizing information so as to facilitate making the recommended inferences.
5. It is a medium of human expression, i.e., a language in which we say things about the world.

The paper stresses that describing these five roles would give a clear framework for a representation and it could capture the 'mindset' by which it was created.

[16] Davis, R., Shrobe, H., and Szolovits, P., 1993, "What is a Knowledge Representation?" *AI Magazine*, 14(1):17-33.

It also suggests going back to the root of KR, which is the capturing and representing of the richness of the natural world. According to the authors too much attention is put on the computational side of KR and the data structures to be used. They feel that this stifles the development of KR.

It is important to keep computational efficiency in mind as expressed in role 4, but the core reasoning in the development of the representation is of far more importance than the data structures and calculations. It has been shown that the view of intelligent reasoning determines to a high degree what kind of representation is sought after. This also shows that certain representations, although good in a certain field, may not be adequate in another.

2.2.4 Knowledge Base

As can be seen in Figure 3 a knowledge base, capturing the knowledge of a trader in his field of expertise or his own personal experience on the effects of major events happening in the world on the market, will also be required in the interpretation phase of the decision-making process. Niches et al (1991)[17] mention in their paper the need for a knowledge base for knowledge sharing and the impediments faced when developing such a knowledge base.

The impediments mentioned are as follows:

- Heterogeneous Representations: There are a variety of representations. This is unavoidable since a representation good for one subject may not be good enough to capture another subject. Currently this is solved through manual translation.
- Dialects within Language Families: Even within the same language of representation, it is still difficult to communicate with dialects. Dialects may contain unknown formalisms or have a different meaning for an encoding that the main language has. This could be solved by standardizing representations.
- Lack of Communication Conventions: It is not necessary to merge knowledge bases, as long as they can communicate with each other. For this an agreed-on protocol must exist to facilitate communication between different knowledge bases.

[17] Neches, R., Fikes, R., Finin, T., Gruber, T., Patil, R., Senator, T., and Swartout, W.R., 1991, "Enabling Technology For Knowledge Sharing," *AI Magazine*, Volume 12, No. 3.

- Model Mismatches at the Knowledge Level: Finally, as the abovementioned problems are solved, you'll still have problems with shared vocabulary and domain terminology. The same words within one domain could have a different meaning in another domain.

2.3 Related Works

In this section related works of fellow researchers will be quoted to give an idea of possible solutions for the opted problem in this thesis. In surveying the literature, two distinct foci were seen, one on financial markets and the other on KR.

The works of various authors presented here are divided into three approaches. First there will be a financial approach, where the entire problem will be solved in a quantitative manner. The second approach will be the KR approach where studies are presented on how to represent a text in such a manner a computer is able to process it for further analysis, which makes it a more qualitative approach. And finally research projects will be cited where the two approaches are combined, which will be called a dichotomous approach. Possible representations will be given of news messages which are then used for analysis in the development of asset prices.

2.3.1 Financial approach

This approach is one often taken in the financial arena. It entails quantifying all the information available and then trying to find a pattern in all the data. It has come to the attention of many researchers that news has an effect on market prices. Many have developed various techniques in utilizing this aspect for predicting asset price movement.

There are mainly two classical techniques employed by economists for asset price analysis: fundamental analysis and technical analysis. Fundamental analysis is based on studying the fundamentals of companies, like balance sheets, price/earning ratio and cash flows. Its success is mainly due to the success of the portfolio theory by Markowitz and the CAPM theory, with their beta coefficient development. Technical analysis has several

branches, but is well defined by the chart analysis. This technique recognizes that today and future prices are in some sense linked to the past prices, which is contrary to the weak form of the EMH.

Blasco et al (2005)[18] analyse the asymmetric effect of news on Spanish stocks, meaning that the authors believe that stock prices are significantly less affected by good news than bad news, which is also suggested by herding behaviour theories. Their methodology relies mainly on regression analysis, ordinary least squares (OLS) in particular, on returns with volatility and trading volume. As a side note, they also employ dummy variables to filter out calendar effects, dummies for day-of-the-week effects and monthly effects, so only the effect of the announcement itself will be captured. Kaminsky and Schmukler (1999)[19] use similar techniques in analysing what type of news moves the markets during the Asian crisis in 1997-1998.

Cho and Engle (2000)[20] also study the asymmetric effects of news. They focus mainly on the beta, investigating whether or not a beta increases with bad news and decreases with good news, just as volatility does. Again, their methodology relies on regression. They introduce a double-beta model with EGARCH variance. The double-beta model was specified, because they wanted to distinguish two effects in the individual excess returns of the firms, the market factor effect and its idiosyncratic effect.

An IMF Working Paper, Funke and Matsuda (2002)[3], which was also mentioned earlier analyses the link between macroeconomic news and stock prices. They hypothesize that the impact is dependent on the type of news, type of stock, the state of the economy, and the integration of the country into the world economy. They also suggest that the impact occurs with a short period of time.

[18] Blasco, N., Corredor, P., Del Rio, C., Santamaria, R., 2005, "Bad news and Dow Jones make the Spanish stocks go round," *European Journal of Operational Research* 163, pp. 253-75.
[19] Kaminsky, G.L., and Schmukler, S.L., 1999, "What triggers market jitters: A Chronicle of the Asian Crisis," *International Finance Discussion Papers*, Number 634.
[20] Cho, Y-H., and Engle, R.F., 2000, "*Time-Varying Betas and Asymmetric Effects of News: Empirical Analysis of Blue Chip Stocks.*"

The news in the abovementioned studies is mainly implicit. The following papers model news and its impact a bit differently and more explicit. Mitchell and Mulherin (1994)[21] model news as number of news announcements and their effect on securities market activity, trading volume and market return in particular. The observed relation between news and market activity was, however, not particularly strong and they did not explain the day-of-the-week seasonality's. Their conclusion on their analysis of the Dow Jones confirmed the difficulty of linking volume and volatility to observed measures of information.

Patell and Wolfson (1982)[22] examine a completely different aspect of news. Their study examines firms' behaviour with respect to the systematic intraday timing of earnings and dividend announcements. Their hypothesis is that good news is more likely to be released when the security markets are open while bad news appears more frequently after the close of trading. Their hypothesis is accepted and they ascribe their results to attempts of management trying to provide a natural no-trading period for the dissemination and evaluation of news releases and to reduce the public exposure of unfavourable events.

2.3.2 KR approach

This approach is from a completely different field of research. Representations of concepts are often expressed by mere words, but the relationship between them or the actual meaning of those words, the semantics, is difficult to capture.

The internet has grown and is still growing rapidly. It contains many objects providing a broad variety of information. The World Wide Web Consortium (W3C), the standardization committee of the WWW, is striving to find a machine-processable representation of the semantics of the sources on the WWW. This is useful in employing the power of automated reasoning.

[21] Mitchell, M.L., and Mulherin, J. H., 1994, "The Impact of Public Information on the Stock Market," *The Journal of Finance*, Vol. 49, No. 3, Papers and Proceedings Fifty-Fourth Annual Meeting of the American Finance Association, Boston, Massachusetts, January 3-5, 1994, 923-50.
[22] Patell, J.M., and Wolfson, M.A., 1982, "Good News, Bad News, and the Intraday Timing of Corporate Disclosures," *The Accounting Review*, Vol. 57, No. 3 (Jul., 1982), pp. 509-27.

W3C, in a collaborative effort with participation from a large number of researchers and industrial partners has developed for this reason a common framework that allows data to be shared and reused across application, enterprise and community boundaries called the *Semantic Web*[23]. It is based on the RDF framework, which integrates a variety of applications using XML for syntax and Uniform Resource Identifiers (URI) for naming. This framework is intended to create a universal medium for information exchange by giving meaning (semantics), in a manner understandable by machines, to the content of documents on the Web.

Harmelen and Fensel (1999)[24] state in their paper for IJCAI '99 Workshop on Intelligent Information Integration that two alternative and complementary strategies are can be used for adding semantics. First, one can enrich information sources declaratively with annotations that provide their semantics in a machine accessible manner. Second, one can write programs (filters, wrappers, extraction programs) that procedurally extract such semantics of Web sources.

There are a number of existing languages for annotating Web sources with semantics:

- HyperText Markup Language (HTML)
- Cascading Style Sheets (CSS)
- eXtensible Markup Language (XML)
- Resource Description Framework (RDF).

HTML is normally used to indicate layout and structure of a document. To add semantics you can use HTML (META)-tags or HTML (SPAN)-tags. (META)-tags intended use is limited to stating global properties that apply to the entire document. As for the (SPAN)-element, it is a generic container of any text element offering a generic mechanism for adding structure to documents.

CSS aim to separate the structure of a document from a specification of the layout of the document. Although the (STYLE)-mechanism was originally intended for layout information, it can also be used and maybe abused for adding semantic information.

[23] http://www.w3.org/2001/sw/
[24] Harmelen, F. van, and Fensel, D., 1999, "*Practical Knowledge representation for the Web*," IJCAI'99 Workshop on Intelligent Information Integration.

XML is used much like HTML, but here the difference is that the user is allowed his own set of markup-tags. These tags can be chosen to reflect the domain specific semantics of the information, rather than merely its lay-out. Although any tags can be used as long as they are properly nested in the document, it is possible to define restrictions on the set of tags that can be used. This is done in a Document Type Definition (DTD), which expresses in a grammar-like formalism which allowed sequences and nestings of tags are allowed in a document.

Finally, we have RDF. In XML structure and semantics are interwoven, whereas RDF provides a means for adding semantics to a document without making any assumptions about the structure of the document. The data model of RDF provides three object types:

- A *resource* is an entity that can be referred to by an address at the WWW (i.e. by an URL). Resources are the elements that are described by RDF statements.
- A *property* defines a binary relation between resources a/or atomic values provided by primitive data type definitions in XML.
- A *statement* specifies for a resource a value for a property. That is, statements provide the actual characterizations of the Web documents.

Another format, which is based on RDF and is widely used, is RSS[25]. RSS is either Rich Site Summary (RSS 0.9x), RDF Site Summary (RSS 0.9 and 1.0) or Really Simple Syndication (RSS 2.x). Whatever the abbreviation stands for, it is used for the same purpose. It provides items containing short descriptions of web content together with a link to the full version of the content. This information is delivered as an XML file called RSS feed, RSS stream, or RSS channel. A program known as a *feed reader* or *aggregator* can check RSS-enabled web pages on behalf of a user and display any updated articles that it finds. RSS is, therefore, used by many major organizations, including Reuters and the Associated Press, after being widely used by weblogs for several years.

[25] Description of RSS is taken from the free online encyclopedia Wikipedia.

Next to KR on the web there are also representation theories which are not completely web-based, like S*emantic Networks*[26]. A semantic network or net is a graphic notation for representing knowledge in patterns of interconnected nodes, which represent concepts and arcs, which represent semantic relations between the concepts. Important *semantic relations*[27] are:

- Meronymy (A is part of B)
- Holonymy (B has A as a part of itself)
- Hyponymy (or troponymy) (A is subordinate of B; A is kind of B)
- Hypernymy (A is superordinate of B)
- Synonymy (A denotes the same as B)
- Antonymy (A denotes the opposite of B)

Semantic Nets were first developed for artificial intelligence and machine translation, but earlier versions have long been used in philosophy, psychology and linguistics. The term 'semantic network' as it is used now might best be thought of as the name for a family of representational schemes rather than a single formalism. The term was first introduced by Ross Quillian in his Ph.D. Thesis (1968)[28], in which it was introduced as a way of talking about the organization of human semantic memory, or memory for word concepts. The idea of a network of associatively linked concepts is claimed by Anderson and Bower (1973)[29] to date all the way back to Aristotle. This network was conceived in the words of Quillian as a "representational format [that would] permit the 'meanings' of words to be stored, so that humanlike use of these meanings is possible"[28].

Sowa (2000)[30] on Semantic Networks describes six of the most common kinds of semantic networks:

- **Definitional networks** emphasize the *subtype* or *is-a* relation between a concept type and a newly defined subtype. The resulting network, also called a *generalization* or *subsumption* hierarchy, supports the rule of inheritance for copying properties defined for

[26] Description of Semantic Networks is taken from the free online encyclopedia Wikipedia.
[27] For a more extensive list of semantic relations see Appendix A, which is a list taken from the free online encyclopedia Wikipedia.
[28] Quillian, M. R., 1968, "Semantic Memory," In Minsky, M. (Ed.) (1968), *Semantic Information Processing*, Cambridge, Mass.: MIT Press, pp. 9, 216.
[29] Anderson, J. R., and Bower, G. H., 1973, "*Human Associative Memory,*" New York, John Wiley and Sons, pp. 9.
[30] Sowa, J.F., 2000, *Knowledge Representation: Logical, Philosophical, and Computational Foundations,* Brooks/Cole Publishing Co., Pacific Grove, CA.

a super-type to all of its subtypes. Since definitions are true by definition, the information in these networks is often assumed to be necessarily true.

- **Assertional networks** are designed to assert propositions. Unlike definitional networks, the information in an assertional network is assumed to be contingently true, unless it is explicitly marked with a modal operator. Some assertional networks have been proposed as models of the conceptual structures underlying natural language semantics.
- **Implicational networks** use implication as the primary relation for connecting nodes. They may be used to represent patterns of beliefs, causality, or inferences.
- **Executable networks** include some mechanism, such as marker passing or attached procedures, which can perform inferences, pass messages, or search for patterns and associations.
- **Learning networks** build or extend their representations by acquiring knowledge from examples. The new knowledge may change the old network by adding and deleting nodes and arcs or by modifying numerical values, called weights, associated with the nodes and arcs.
- **Hybrid networks** combine two or more of the previous techniques, either in a single network or in separate, but closely interacting networks.

With the above descriptions of the different kinds of semantic networks, even the diagrams in this thesis could be considered a semantic net of some sort, whereby the graphical representation tries to capture the semantics and relations between concepts. Having seen the different methods of capturing semantics, an example of an actual application of the use of web-based KR for news can be seen in Mueller (2000)[31].

Mueller (2000)[31] suggests that if news were made available in a computer-understandable form, computers and devices could be made aware of events taking place in the world. He proposes that this awareness could, for example, help navigation systems to avoid traffic jams or scheduling programs to change or cancel a booking, if a natural disaster occurred at the holiday destination. XML documents called NewsForms are then presented that represent key points of 17 types of news events.

[31] Mueller, E.T., 2000, "*Making news understandable to computers*," Signiform, Washington, D.C.

Two XML document types defined for news are mentioned in that paper:

- The News Industry Text Format (NITF) was created by the International Press Telecommunications Council and the Newspaper Association of America. NITF is a text markup language that enables reuse of news stories across print publications, broadcast news, databases, and the web. It incorporates a number of HTML tags for structure, appearance, and linking. It adds tags for metadata or information about a news story. Finally, it also includes tags for marking up entities in the story such as *event, location, money, org, and postaddr*.
- XMLNews was created by WavePhore, a news amalgamator, in order to provide web sites with a feed of news stories gathered from various sources in a standard format. XML consists of XMLNews-Story, which is an easier-to-implement subset of NITF for marking up text news stories, and XMLNews-Meta, which provides information about news stories, text or otherwise.

The NewsForm described by Mueller describes 17 types of news events: competitions, deals, earnings reports, economic releases, Fed watching, IPOs, injuries and fatalities, joint ventures, legal events, medical findings, negotiations, new products, management successions, trips and visits, votes, war, and weather reports. The NewsForm document type defines:

- Elements for types of news events such as *InjuryFatality* and *Deal*
- Child elements such as the *Cause* of an *InjuryFatality* and the *Target* of a *Deal*, and
- Standard values such as *Earthquake* and *Fire* values of the *Cause* of an *InjuryFatality*.

NewsForms can also be created automatically by using a software application, NewsExtract. It uses information extraction techniques to convert text stories into NewsForm, though with some errors and omissions. For the selection of concepts it relies on the use of 59 specialized lexicons. The application consists of the following components:

- Sentence boundary identifier
- Part-of-speech tagger
- Noun group parser
- Entity parser
- Reference resolver
- Pattern-based parser
- Commonsense rules

Mueller suggests that the benefits of computer-understandable news are many. First of all, it allows for tracking and notification. Users can be made aware of events happening that are of special interest to them, be it for online trading or other purposes. Secondly, it facilitates historical studies on news events, such as in this study the impact of news on financial markets. It could also assist in producing informational graphics, by computing a number of statistics and visualizing them. Finally, it could allow calendar and other application programs to be more aware of the world. With the awareness of news, an online calendar could inform the user if an appointment is in a building that has just been evacuated due to an accident or suggest alternative means of transportation in case of a subway outage.

2.3.3 Dichotomous approach

In the previous two approaches the focus has either been on the financial market or the representation of news. In this section descriptions of studies will follow that try to find a suitable representation for news as to be able to use it for an analysis of the market and its price development process. In the financial approach an initial attempt was made, but they often disregard the content of a news announcement or mention news only in an implicit manner. In the KR approach representations were discussed and here even more representations will be presented in an attempt to capture the content of a news message.

Baestaens and van den Bergh (1996)[32] suggest in their analysis of public information effects on the DEM/USD swap rate, a routine to treat single line text of headlines with a view to predicting the return series in operational time. The headlines used were retrieved from Money Market Headline News.

The routine to treat the headlines is as follows. Firstly, for all news flashes in the training set the frequency of every single word was counted. Highly frequent but uninformative words were filtered out. Moreover, they also standardized words to increase their frequency of appearance. In this manner 15,474 words were identified. Words with a frequency lower than 100 were eliminated. The assumption made here was that low frequency

[32] Baestaens, D.J.E., and Van den Bergh, W.M., 1996, "*Public information Effects on the DEM/USD swap rate: An intraday analysis in operational time*," Rotterdam Institute for Business Economic Studies, R 9602/F, Erasmus University Rotterdam.

words were to relate to rather unique, exceptional and unsystematic news items. With the identified words the original news sentences were reconstructed. To increase the similarity of sentences they were first restricted to a maximum of four words. Then, the words in each sentence were sorted following the word frequency in descending order. This procedure resulted in the formation of clusters of possible similar sentences. These newly constructed sentences were then paired with actual swap returns. Both linear and non-linear models were specified to test whether the effective news component is a statistically useful explanatory variable. The result shows that it did not improve the forecast error, but appeared to capture a part of the dependencies within the swap return unaccounted for by the more traditional technical variables.

Jacob and Rau (1990)[33], although also searching for a way to capture news in a type of representation, use a different approach. They analyse a prototype system called SCISOR, which stands for System for Conceptual Information Summarization, Organization, and Retrieval.

This system performs text analysis and question answering within a constrained domain. It performs the following tasks: lexical analysis, finding story structures, topic determination, natural language analysis and the storage/retrieval of conceptual representations into and out of a knowledge base. This system combines artificial intelligence (AI) methods, especially natural language processing, KR, and information retrieval (IR) techniques, with more robust but superficial methods, such as lexical analysis and word-based text search.

The difference in approach is its constraint to stories only about takeovers. The conceptual representation chosen is something like *Corporation-takeover-offer* and can be further split into *Offerer, Offeree* and *Offered*. When a question is inputted into the system, it uses a conceptual graph-matching algorithm that ranks and compares the representation of the question to representations of stories.

[33] Jacobs, P.S., and Rau, L.F., 1990, "SCISOR: Extracting Information from On-line News. An Object-Oriented Relational Database," *Communications of the ACM*, Vol. 33, No. 11, pp 87-97.

A study of SCISOR against a traditional IR system was performed which may have been futile and difficult, because this system works only in constrained domains whereas traditional IR are often tested on arbitrary, unconstrained texts. Another reason is that this system performs many tasks other than document retrieval, such as extracting information from stories and directly answering users' questions. But seeing that this program accurately processed financial news, selected items of interest, provided a user interface that allows easy access to results, and extracted important information in a structured form, it is this sort of text-based conceptual information system that will be at the heart of intelligent systems technology.

Bunningen (2004)[34] examines the representation problem in a similar manner as SCISOR does. He also focuses his work in a constrained domain, mainly the pharmaceutical industry. The extraction procedure was based on extracting features from news articles into a template. This extraction was performed syntactically and not semantically as semantic analysis was not deemed attractive at that time. An important aspect for the extraction process is the domain knowledge (In Appendix B types of domain knowledge are listed). The representation that was chosen for his experiment was the Term Frequency Inverse Document Frequency (TFIDF), but instead of applying it to terms it was applied to features. With the TFIDF a model for text classification was specified by labelling the feature vectors with either surge, plunge or not relevant. Support Vector Machines (SVM) was then used to solve the classification problem.

Hariharan (2004)[35] also uses TFIDF as a representation for news stories and SVM for recognizing patterns in the training data. The difference is the purpose of the research. Hariharan in his thesis is specifically analysing automatic trading agents. He suggests a trading strategy using online news as an external input. The performance of the trading agents was evaluated using a standard academic and industry metric, the Sharpe Ratio. The results demonstrate the power of online news in business.

[34] Bunningen, A.H. van, 2000, "*Augmented Trading*," Master Thesis, University of Twente, Enschede, The Netherlands.
[35] Hariharan, G., 2004, "*News Mining Agent for Automated Stock Trading*," The University of Texas at Austin, USA.

2.4 Summary

In this chapter the result of an extensive literature survey was reported. Three approaches: financial, KR and dichotomous, were discussed. Taking into account the related works on the problem a more detailed representation can be given then shown earlier in Figure 1.

A possible solution to the problem can be more specifically defined. Both areas, news and the market, can be modelled or represented in a certain way and those representations are then matched. In Figure 4 it is shown that what can be considered as news is the entire news article, a summary or abstract thereof, merely the headline, the type of news or the quantity. A representation or interpretation of news could then be e.g. the frequency of words or word groups, a vector containing features of an article, headlines, fingerprints or an association matrix. On the other hand, what can be considered as the market are the stocks, indices, options, swaps, futures, forwards or currency. Representations of these are the prices, returns, trading volume, volatility or just the going up or down of any asset price. Testing the relationship between these matched pairs could clarify or describe the relationship that exists between news and the market. This in a way could give an idea of the effect of news on the market.

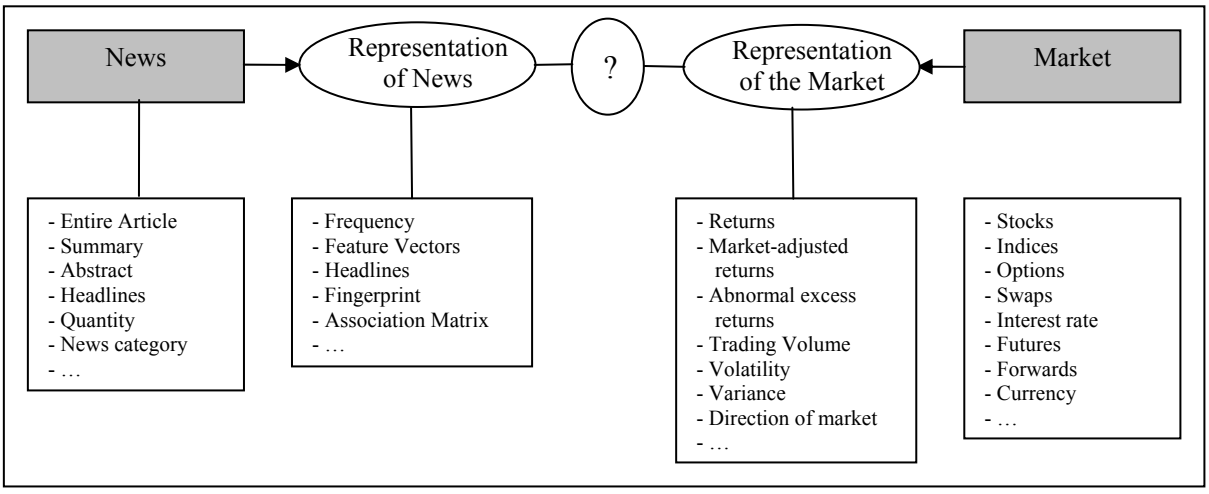


Figure 4: A different representation of the problem.

Another benefit of describing the problem in this manner is its power to predict. To predict market movement new news messages will be converted into the representation chosen earlier for training and then compared to existing representations of news in the training data. This procedure (shown in Figure 5) reduces the complex problem into a much simpler classification problem. Simple in understanding its basic concepts and fundamental ideas, but the application of this procedure can still be quite complex seeing that it entails many choices that have to be made.

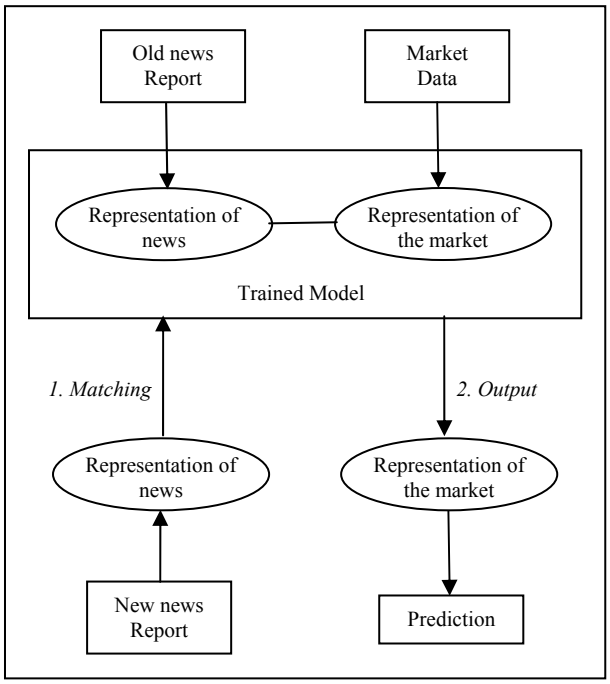


Figure 5: Procedure for Prediction.

With all this in mind, the next chapter will describe the methodology behind the experimentation utilised in this thesis. It will contain the choices made with regard to the representations both for the market and for news. The algorithm which will be used for modelling and prediction purposes will also be described in the next chapter.

This thesis experiment is therefore setup to see if the relationship between news and the market can be quantified and thereby getting an edge by predicting market movements, so as to assist in financial decision making.

3 METHODOLOGY

"If we knew what we were doing, it wouldn't be called research, would it?"

--Albert Einstein (1879-1955)

"The important thing in science is not so much to obtain new facts as to discover new ways of thinking about them."

--Sir William Bragg (1862 - 1942)

3.1 Introduction

In the previous chapter an account was given on a survey done on literature concerning the problem in this thesis. After close examination of various papers written by fellow researchers a choice was made to answer the questions proposed in chapter 1 by treating the problem as a text classification problem. Figure 5 shows the procedure in general that is going to be followed in this study.

Although the choice has been made to treat the problem as a classification problem, it still leaves many questions unanswered. First of all, what is the representation going to be for the market? Secondly, what is going to be used as a representation for news reports? Thirdly, after the matching process, which algorithm is going to be used to distinguish the various groups, if they exist? In other words, which technique is going to be used to compare new news messages with?

To all these questions an answer is going to be given in this chapter. Starting first by showing what is going to be used to model the market, and then followed by a description of the representation chosen for news. This chapter will end with an account of the algorithm

used for modelling the proposed problem. In the next chapter the experimental setup and how the experiment was actually performed will be described.

3.2 Representation of the Market

Since asset prices are mostly established electronically these days, a lot of data is being stored electronically in financial databases. The financial market has various types of securities, like stock prices, options, futures, warrants and swaps. For each security various kinds of prices are kept. What are often freely available are the open and closing prices, daily highest and lowest price and trading volume.

From these prices various other information can be calculated, for instance returns, volatility and variance. As the problem is being treated as a text classification problem, classes or labels need to be found for the news messages by using market data. In Hariharan (2004)[35] two methods of labelling are mentioned. There is manual labelling, whereby an expert is consulted and news reports are labelled by hand according to the market sentiment they are presumably going to produce. This requires, however, the assistance of an expert and the selection procedure to find such an expert would take to much time and is beyond the scope of this thesis. Hariharan suggests that the errors found in hand labelling could be compensated by a large number of training instances, which led him to another type of labelling, automatic labelling.

But to do automatic labelling a performance measure is needed. In section 2.2.1 about the EMH the abnormal return was found to be a good reflection of the performance of a security. In that section the *abnormal return* is calculated in the following manner:

$$\textit{Abnormal return} = \textit{actual return} - \alpha + \beta \text{ x } \textit{return on market index}$$

where α states how much on average the stock price changed when the market index was unchanged and β tells us how much extra the asset price moved for each 1 percent change in the market index.

The abnormal returns can then be divided into three different categories: rise, plunge and not relevant, signifying the market sentiment of that day. The division of the three classes will be determined by the size of the not relevant class. For research purposes the parameter gamma (γ) will be used to adjust the size to determine a suitable size for the three categories. The news stories will then be labelled according to the sentiment of that day. For empirical purposes the raw returns without market correction will also be used for labelling following the same procedure. Doing both procedures may shed more light on the effect of news on the market.

Another way to label news announcements automatically would be to use the trading volume. This is an idea taken from Mitchell and Mulherin (1994)[21] where they study the effect of the number of announcements to trading volume. Although the trading volume cannot help in answering the question proposed in this thesis if the news can predict market movements, it can however give an idea of the possible effects news has on its participants.

For labelling purposes the daily average volume traded will be calculated over the complete period and volumes above or below the average will be considered as exceptional trading behaviour. Another method for modelling, which may deserve testing would be to only consider exceptional trading behaviour when the trading volume rises or drops above or below one standard deviation from the average volume. The choice whether which method will be used will depend on the number of training instances left for further testing. Although this may not be a very scientific approach in choosing the best method, the volume labelling is only going to be used as an indication. There are many more complex ways to model volumes, but again seeing that the volume modelling will only be used as an indication of market sentiment and the behaviour of market participants no further complexity is deemed necessary for modelling.

3.3 Representation of News Messages[36]

For the representation of news announcements the solution for KR of Collexis® was chosen. The choice for utilising the Collexis® software is primarily because of the familiarity with the Collexis® technology and its success in the information retrieval community. In this section background information on the technology, why it was developed and a description of Collexis® KR, their Fingerprinting Concept, will be given.

3.3.1 Background on Collexis®

As mentioned earlier the availability of information has increased tremendously in the last few decades. Many search engines on the WWW often only search for single information items. In a knowledge intensive environment that is no longer sufficient. Current retrieval methods range from data retrieval (results are single information items, like a document/article/URL) to more advanced information retrieval (results presented by related information items like documents but also experts).

The approach Collexis® technology takes is new, that of knowledge retrieval. Besides having the capabilities to retrieve information, it is also able to discover relationships between elements of different information items (via clustering and/or aggregation) and thus uncover implicit information. The retrieval results obtained have both been impressive in terms of quality and performance.

3.3.2 Why Collexis® was developed

Collexis® was developed to provide ultimate quality and performance when unlocking large amounts of electronic data: To achieve this quality and performance, the technology is capable of providing results to a search in even millions of documents instantly. The performance is also increased by being able to search both structured and non-structured data.

[36] Information adapted from the Collexis® Product Overview.

High quality is maintained by not only identifying documents but also relevant experts and organizations.

Two quality measures widely used in information retrieval are recall and precision[37]. Recall is defined as the proportion of relevant information identified and precision is defined as the proportion of selected information that is relevant. To achieve a high recall whereby relevant documents are not excluded, even when narrowing a search, synonyms in a thesaurus are included in the search resulting in more relevant results being found. The search terms/keywords can also be expanded with more specific terms/keywords defined in a thesaurus. To improve precision on the other hand, amongst others the homonym disambiguation module which can distinct between different word senses of the same word is used in identifying only relevant documents and/or experts. It also utilises a threshold which makes it possible to indicate how many relevant results have to be retrieved.

3.3.3 The Idea behind Collexis®

Collexis® is based on the principle of Fingerprinting. The idea is that a fingerprint, although small, is a unique representation of a person, so a fingerprint of a piece of information (such as a publication in MS-Word, an e-mail, a PowerPoint presentation, a selected text on a website) is also a unique representation of that piece of information. Fingerprints are created of each piece of information using the knowledge residing in a thesaurus.

Collexis® creates a fingerprint out of any text such as an article, a competence sheet, a project description, an e-mail or a web page. The fingerprint can also be a multi-facet fingerprint, which means it is composed of a number of sub-fingerprints each of them using a different thesaurus. This multi-facet fingerprint can in addition include a free text fingerprint, i.e. a fingerprint that is based on the individual words in the text (as opposed to those made up of concepts defined in a thesaurus).

[37] Eck, N.J. van, 2005, "*Towards Automatic Knowledge Discovery from Scientific Literature, Computer Based Tools for Supporting Scientific Research,"* Maste Thesis, Erasmus University Rotterdam, The Netherlands.

Collexis® not only creates fingerprints from all content, but also from the search information. This information can be a few words, a sentence, but also a complete document. By comparing the search and content fingerprints, Collexis® provides only highly relevant information within milliseconds.

Figure 6: Sample Collexis® Fingerprint.

A Collexis® fingerprint is a profile of a text, person or organization in concepts (keywords) each with its relative weight (See Figure 6 for a sample fingerprint of the concept *Malaria*). The concepts in a fingerprint are selected by comparing the data set with a thesaurus which contains all relevant concepts. A complex set of algorithms determines the selection and weight of each concept in the fingerprint. They are created from both the search texts as the texts in the database(s). The idea behind Collexis® is depicted in Figure 7.

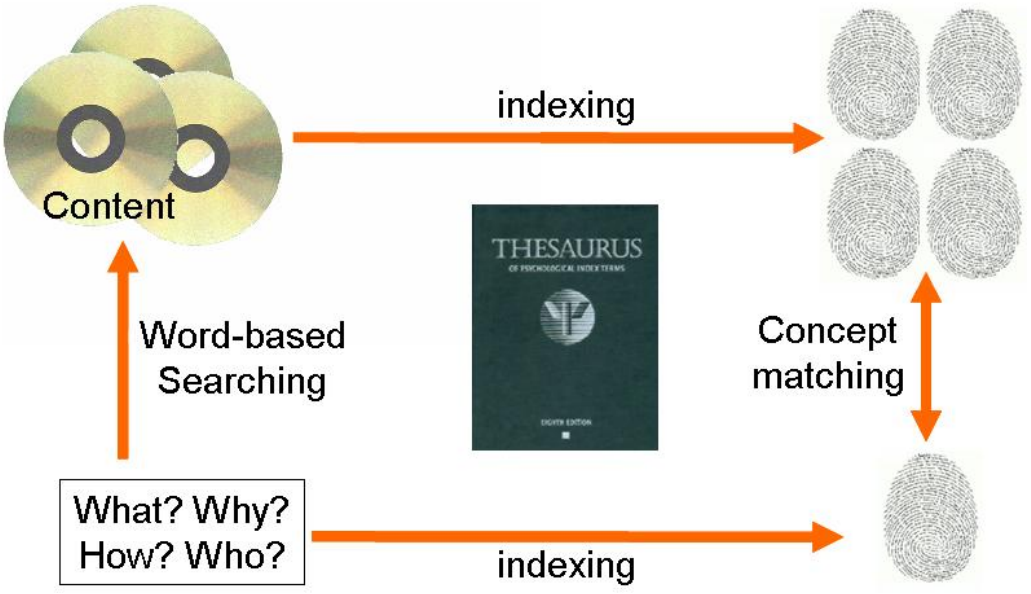


Figure 7: The Idea behind Collexis®.

3.3.4 Collexis® fingerprint

As can be seen in Figure 6 a fingerprint is a list of concepts with its relative weights. How are concepts selected? How are the weights determined? Because Collexis® is a commercial software package only an overview of fingerprinting is given.

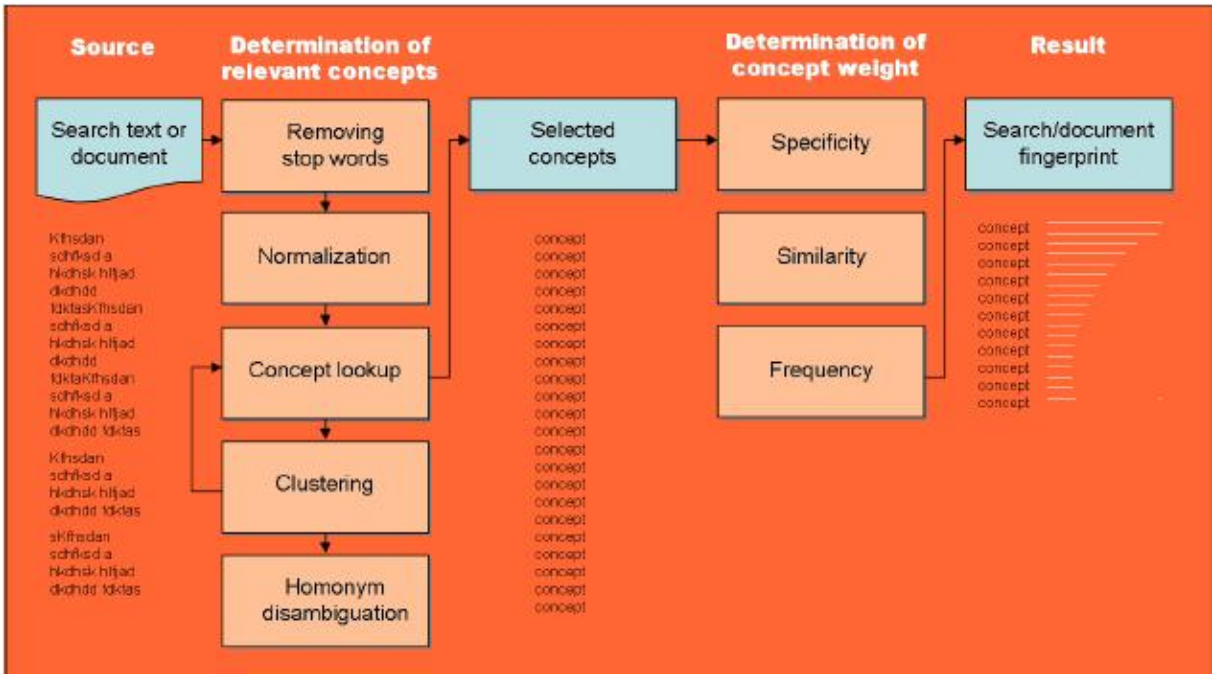


Figure 8: Abstraction Process.

The abstraction process or the fingerprinting process itself is shown in Figure 8. First the stop words are removed from the source text, then the words are normalised. Other procedures are also performed and the concepts are further selected by use of a thesaurus based on specificity, similarity and frequency, resulting in a vector of concepts with corresponding weights.

For matching purposes both the fingerprint of the query as well as the source document can be seen as vectors in a vector space. The distance between those two vectors can be seen as a measure of how relevant that document is to the query[38]. In Figure 9 this is illustrated by a Vector Space Model based on three concepts. In the same manner as matching query fingerprint to source document fingerprint, new news fingerprints can be matched to

[38] Appendix C contains an adapted version of the Collexis API Documentation showing a list of algorithms used by for matching a query to a record.

news fingerprints from the training set. New news fingerprints will then get the same label as the label of the nearest news fingerprints or group of news fingerprints.

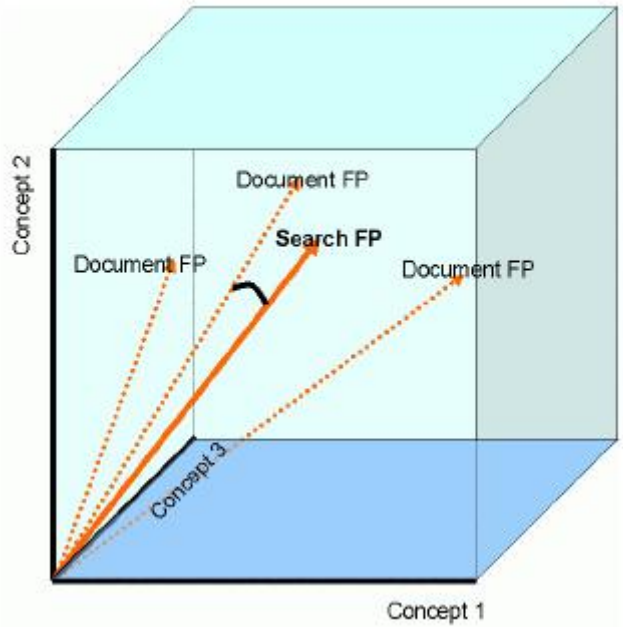


Figure 9: Vector Space Model based on 3 Concepts.

3.5 Support Vector Machines[39] [40] [41]

To solve the classification problem, SVM will be utilised, partially motivated by the use in Hariharan (2004)[35]. SVM are based on the concept of decision planes that define decision boundaries. The decision plane separates a set of objects belonging to different classes. In Figure 10 an example of a linear decision plane is shown. Most classification tasks are, however, not that simple. Many times a more complex structure needs to be found to make an optimal separation (See Figure 11 for non-linear separation).

[39] Description of Support Vector Machine is taken from the free online encyclopedia Wikipedia.
[40] Hastie, T., Tibshirani, R., and Friedman, J., 2001, *The Elements of Statistical Learning; Data Mining, Inference, and Prediction,* Springer-Verlag, New York.
[41] http://www.statsoft.com/textbook/stsvm.html

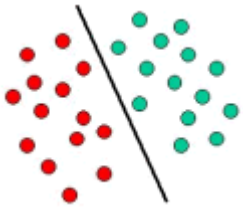


Figure 10: Classic Example of linear classifier.

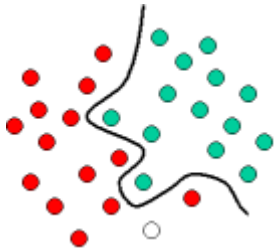


Figure 11: Classic example of non-linear classifier.

Basic SVM creates an optimal separating hyperplane that lies in a transformed input space. In Figure 12 you can see the original objects (left side of the schematic) mapped, i.e., rearranged, using a set of mathematical functions, known as kernels. This process of rearranging the objects is called mapping or transformation. Note that in the new setting, the objects are linearly separable and, thus, only an optimal line has to be found to separate the two classes. An optimal separating hyperplane separates two classes and maximizes the distance to the closest point from either class. By maximizing the margin between two classes on training data, it often leads to better classification performance on test data.

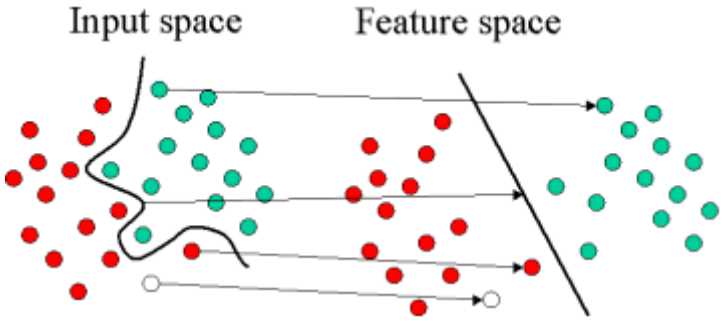


Figure 12: Basic idea of Support Vector Machines.

The use of the maximum-margin hyperplane is motivated by Vapnik Chervonenkis theory (also known as VC theory) which was developed by Vladimir Vapnik and Alexey Chervonenkis. The theory is a form of computational learning theory, which attempts to explain the learning process from a statistical point of view. This theory lead to a probabilistic test error bound which is minimized when the margin is maximized (See Vapnik (1996)[42]). Due to the very large slack associated with the bounds, the utility of this theoretical analysis is sometimes questioned. The parameters of the maximum-margin hyperplane are derived by

[42] Reference taken from Hastie, Tibshirani and Friendman (2001). Vapnik, V., 1996, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York.

solving a quadratic programming (QP) optimization problem. There exist several specialized algorithms for quickly solving the QP problem that arises from SVMs.

The original optimal hyperplane algorithm proposed by Vladimir Vapnik in 1963 was a linear classifier. However, in 1992, Bernhard Boser, Isabelle Guyon and Vapnik suggested a way to create non-linear classifiers by applying the kernel trick (originally proposed by Aizerman) to maximum-margin hyperplanes. The resulting algorithm is formally similar, except that every dot product is replaced by a non-linear kernel function. This causes the linear algorithm to operate in a different space and fit the maximum margin hyperplane in that space. It is a space of constructed features, a non-linear map from the original input space, usually of much higher dimensionality than the original input space.

Three popular choices for the kernel are mentioned in SVM literature: d^{th} Degree polynomial, Radial basis and Neural Networks. If the kernel used is a radial basis function, the corresponding feature space is a Hilbert space of infinite dimension. Maximum margin classifiers are well regularized, so the infinite dimension does not spoil the results.

In 1995, Corinna Cortes and Vapnik suggested a modified maximum margin idea that allows for mislabelled examples. If there exists no hyperplane that can split examples, the Soft Margin method will choose a hyperplane that splits the examples as cleanly as possible, while still maximizing the distance to the nearest cleanly split examples. This work popularized the expression SVM. The SVM was popularized in the machine learning community by Bernhard Schölkopf[43] in his 1997 PhD thesis, which compared it to other methods.

[43] Reference taken from Wikipedia. Schölkopf, B., 1997, *Support vector learning*, GMD-Berichte No. 287, GMD-Forschungszentrum Informationstechnik.

4 EXPERIMENTAL SETUP

"It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts."

--Sir Arthur Conan Doyle (1859 - 1930)

"You can use all the quantitative data you can get, but you still have to distrust it and use your own intelligence and judgment."

--Alvin Toffler

4.1 Introduction

In the previous chapter both the representations of the market and news were expounded. A description of the algorithm going to be used for modelling and prediction was also given at the close of the chapter.

In this chapter the experimental setup will be discussed. First a description of the dataset will be given, both the data for the market as well as for news. Then, an account will be given on the pre-processing of the data, which are the necessary procedures for labelling, so data can be used as input for the experiment going to be performed. The chapter will be concluded by a description of the steps taken to perform the experiment itself, which will consist of both the training and testing of the model. The results will be given in the next chapter.

4.2 Raw Dataset

In this section a description of both the data used for the market as well as news are given. The processing of the data to make it suitable and useable will be given in the next section.

4.2.1 Prices and Trading Volume

For the representation of the market, stock prices were chosen, because of earlier experience with modelling stock prices[44]. The open-close prices, high-low prices and trading volume were retrieved from both DataStream© as well as Yahoo! Finance. The choice was made to use the prices of Exxon Mobil Corp. because the company primarily deals with Oil and Gas. It was deemed interesting to use this company, because intuitively news on Oil and/or Gas often results in market volatility. Apart from this reason, the choice was arbitrary. It was, however, a conscious choice to only model the market by one company. It may be of interest to use more companies in further research or even not limit oneself only to stock prices, but that will considered beyond the scope of this thesis.

First a short history taken from their website[45] will be given on the company used. Exxon and Mobil trace their roots back to the late 19th century, when American industry was booming. John D. Rockefeller acquired a diversity of petroleum interests during that period and, in 1882, organized them under the Standard Oil Trust. That same year marked the incorporation of two refining and marketing organizations: Standard Oil Co. of New Jersey and Standard Oil Co. of New York. "Jersey Standard" and "Socony," as they were commonly known were the chief predecessor companies of Exxon and Mobil, respectively. In 1911, the U.S. Supreme Court ordered the dissolution of the Standard Oil Trust, resulting in the spin-off of 34 companies, including Jersey Standard and Socony. In the same year, the nation's kerosene output was eclipsed for the first time by a formerly discarded by-product - gasoline.

[44] Tan, F.H., 2005, "*Option Pricing, the GARCH-M Approach,"* Bachelor Thesis, Erasmus University Rotterdam, The Netherlands
[45] http://exxonmobil.com/Corporate/About/History/Corp_A_History.asp

The growing automotive market ultimately inspired the product trademark Mobiloil, registered by Socony in 1920.

Mobil Chemical Company was established in 1960. Exxon Chemical Company became a worldwide organization in 1965. The two chemical companies combined their operations within Exxon Mobil Chemical. In 1955, Socony-Vacuum became Socony Mobil Oil Co. and, in 1966, simply Mobil Oil Corp. A decade later, a newly incorporated Mobil Corporation embraced Mobil Oil as a wholly owned subsidiary. Jersey Standard changed its name to Exxon Corporation in 1972 and established Exxon as an uncontested trademark throughout the United States. In other parts of the world, Exxon and its affiliated companies continued to use its long-time Esso trademark and affiliate name.

In 1998, Exxon and Mobil signed a definitive agreement to merge and form a new company called Exxon Mobil Corporation. After shareholder and regulatory approvals, the merger was completed November 30, 1999.

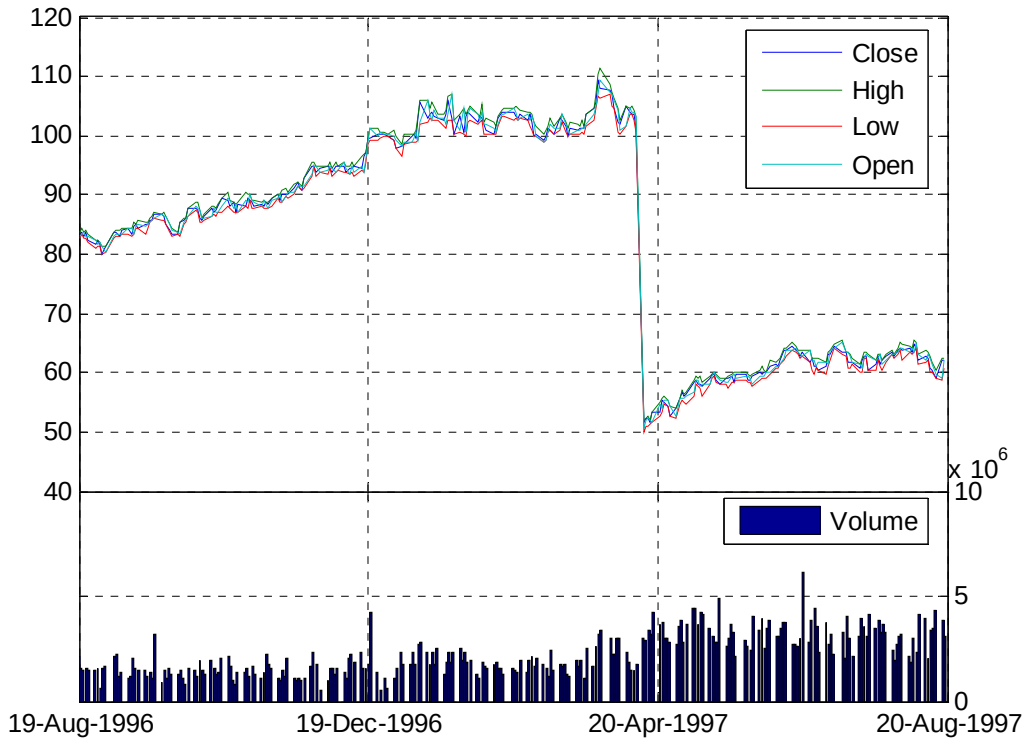


Figure 13: Open-Close Prices, High-Low Prices and Trading Volume of Exxon Mobil Corp.

The charts of the retrieved prices and volume of Exxon Mobil Corp are given in Figure 13 for the period of the Corpus of News being used for this experiment to give an impression of the development of the stock prices.

A sudden drop in the prices is observed around April 1997. In the press releases it is found to be due to a stock split. To get a closer look at the development of the prices the data is split into before and after the stock split as can be seen in Figure 14 and Figure 15.

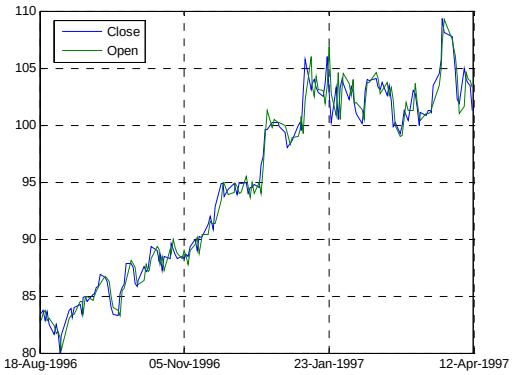


Figure 14: Open-Close Prices of Exxon Mobil Corp. from 19-08-1996 to 11-04-1997.

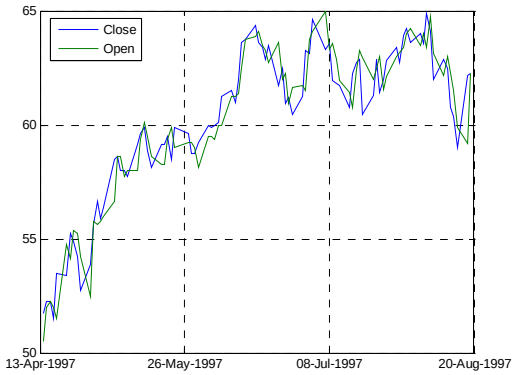


Figure 15: Open-Close Prices of Exxon Mobil Corp. from 14-04-1997 to 20-08-1997.

In the previous chapter on the representation of the market it was also mentioned that a market index was necessary to calculate the abnormal returns, which was suggested to be the difference of the actual return deducted by the expected return. As Exxon Mobil Corp is listed in United States of America, the choice was made to use Standard & Poor's US Index 500 (S&P 500)[46] as the benchmark for the market. This index is widely regarded as the best single gauge of the U.S. equities market. This world-renowned index includes a representative sample of 500 leading companies in leading industries of the U.S. economy. Although the S&P 500 focuses on the large-cap segment of the market, with over 80% coverage of U.S. equities, it is also an ideal proxy for the total market. An important note has to be made here; the S&P 500 contains the Exxon Mobil Corp price.

The S&P 500 is maintained by the S&P Index Committee, whose members include Standard and Poor's economists and index analysts. Committee oversight gives investors the benefit of Standard and Poor's depth of experience, research and analytic capabilities. The

[46] The following description of the S&P 500 is taken from Standard and Poor's website: http://www2.standardandpoors.com.

Committee establishes Index Committee Policy used to maintain the indices in an independent and objective manner.

The history of the S&P 500 dates back to 1923, when Standard and Poor's introduced an index covering 233 companies. The index, as it is known today, was introduced in 1957 when it was expanded to include 500 companies (See Appendix D for a list of the current constituents). In Figure 16 the open and close prices of the S&P 500 are charted to give a general idea of the market sentiment in the period under examination.

Figure 16: Open-Close Prices of the S&P US Index 500 from 19-08-1996 to 20-08-1997.

4.2.2 News Announcements

As for the news messages used, the choice was made to use Reuters Corpus, Volume 1[47]. This Corpus was made available by Reuters Ltd for use in research and development of

[47] In Fall of 2004, NIST took over distribution of Reuters Corpus, Volume 1 and any future Reuters Corpora. The Corpus can be requested via http://trec.nist.gov/data/reuters/reuters.html.

natural language processing, information retrieval, and machine learning systems. The Corpus consists of 806,791 XML files in NewsML format. They are distributed in the form of 365 zip files, one per day, over 2 CDs. Approximately 2.3Gb is required for the storage of the uncompressed XML files. Due to seasonal variations the number of stories per day is not constant, but on weekdays there are on average of 2,880 per day and 480 on weekends. It contains news from 20 August 1996 until 19 August 1997. For more details on the Corpus see Appendix E.

The following is the general structure of the XML file used by NewsML, the format used by Reuters to distribute its news messages:

```
<newsitem>
        <title> </title>
        <headline> </headline>
        <dateline> </dateline>
        <text> </text>
        <copyright> </copyright>
        <metadata>
                <codes class="bip:countries:1.0"> </codes>
                <codes class="bip:industries:1.0"> </codes>
                <codes class="bip:topics:1.0"> </codes>
        </metadata>
</newsitem>
```

For a more detailed structure an example of a NewsML file is given in Appendix F.

4.3 Pre-processing

In this section a description will be given on the work done on the acquired dataset to make it usable for experimentation. First the process of creating the labels will be given by using stock prices and trading volume. Then an account of the selection procedure of the news messages from the large corpus will be given to both reduce the dataset and do an initial selection of messages. A description of the fingerprinting process will also be given in this

section, concluded by describing the created dataset which will serve as input for modelling and prediction.

4.3.1 Labelling by Returns

As mentioned earlier the abnormal returns will be used for labelling the news messages. Once again, the abnormal return is the difference between the actual return and the expected return. The expected return can then be defined by using a factor model dependent on the market index. This relationship can be found simply by regressing the actual stock price returns on the returns of the market index, using the market index as an indication for the expectation of market participants on the stock price returns or more generally the market sentiment.

But before the regression can be done, the returns need to be calculated from the prices in the dataset. The simple periodic returns will be calculated of both the Exxon Mobil Corp. prices and the S&P 500 prices. For the regression procedure to find the parameters for the market model going to be used additional data consisting of open and close prices was retrieved. To estimate the sensitivity of the company stock price to the market three periods of data were obtained. The first period (02-01-1991 to 29-12-1995) is before the period being used in this experiment (20-08-1996 to 19-08-1997), the second period (02-01-1996 to 31-12-1997) contains the period and the third (02-01-1998 to 18-07-2001) is after the period.

A shorter time horizon is more preferable as was seen in the literature survey, intraday prices would have been the best choice, but seeing that historic intraday prices are either expensive or unavailable only daily open and close prices were obtained. Another problem which was faced was the lack of a timestamp on the Reuters Corpus, Volume 1. To still be able to give an indication of the effect of time horizon, it was decided to use open-to-close returns and open-to-open returns testing which would give better predictions for the news messages of a certain day. This in turn could give an idea on the impact of the news reports used. As a timestamp is not available for the announcements this method is not truly reliable, but the results may give an indication of the time horizon effect which has already been established in previous research.

With the abovementioned choice the regression for the market model was performed for the three periods both on open-to-close returns as well as open-to-open returns. For the sake of completeness the regression was also performed on the entire period. The results are given in Table 1 and Table 2.

	Open-to-Close					
	Variable	**Coefficient**	**t-statistic**	**t-probability**	**R-squared**	**Num. Obs.**
Complete Period	**α**	-0,000246	-1,043081	0,297006	0,093	2661
02/01/91-18/07/01	**β**	0,400212	16,511665	0		
Period 1	**α**	-0,000342	-1,316988	0,188082	0,1152	1263
02/01/91-29/12/95	**β**	0,514484	12,811997	0		
Period 2	**α**	0,000512	0,989011	0,323132	0,3127	506
02/01/96-31/12/97	**β**	0,808724	15,143694	0		
Period 3	**α**	-0,000293	-0,5758	0,564896	0,0387	892
02/01/98-18/07/01	**β**	0,233615	5,989737	0		

Table 1: Market Model parameter estimates using Open-to-Close returns.

	Open-to-Open					
	Variable	**Coefficient**	**t-statistic**	**t-probability**	**R-squared**	**Num. Obs.**
Complete Period	**α**	0,000063	0,269645	0,787455	0,1107	2661
02/01/91-18/07/01	**β**	0,49377	18,193167	0		
Period 1	**α**	-0,00001	-0,039429	0,968555	0,182	1263
02/01/91-29/12/95	**β**	0,705407	16,749683	0		
Period 2	**α**	-0,000621	-1,126312	0,260569	0,286	506
02/01/96-31/12/97	**β**	0,961012	14,207683	0		
Period 3	**α**	0,00029	0,571348	0,567908	0,0506	892
02/01/98-18/07/01	**β**	0,297548	6,889612	0		

Table 2: Market Model parameter estimates using Open-to-Open returns.

Looking at the explanatory power of the models estimated, Period 2 model may be the best model using either Open-to-Close or Open-to-Open returns. The constant **α** is in all instances not significantly different from zero. For this experiment the model estimated by Period 2 will therefore be used. It must be noted, however, that although the highest R^2 was used, R^2 was still very low. This low value would indicate a very weak relationship between the company returns and the market returns. The experiment will proceed, but we should keep in mind that when we see little difference between normal returns and abnormal returns, it could be due to this regression and its low explanatory power.

To get an idea of what the dataset looks like and how different the returns are for the calculated abnormal returns, the charts are shown for each time series separately in Appendix G. For an initial inspection and comparison of the differences between Open-to-Close and Open-to-Open returns take a look at Figure 17 where they are plotted for the normal

unadjusted returns and Figure 18 where they are plotted for the abnormal returns. There seem to be significant differences on visual inspection.

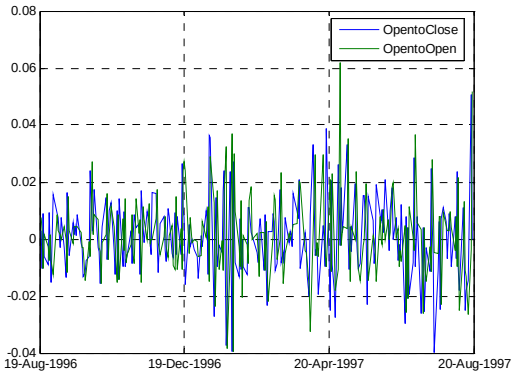


Figure 17: Comparison Open-to-Close and Open-to-Open returns for the unadjusted returns.

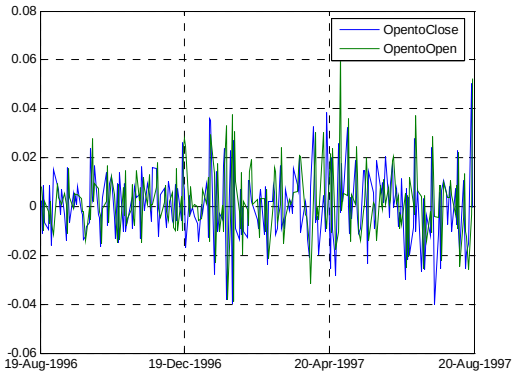


Figure 18: Comparison Open-to-Close and Open-to-Open returns for the abnormal returns.

In Figure 19 and Figure 20 the comparison plots are given of Open-to-Close and Open-to-Open returns respectively for the normal and abnormal returns. In the charts can be seen that abnormal returns are slightly smaller than normal returns, which is expected as the market effect is subtracted from the normal unadjusted returns hoping to only capture company specific effects and cancel out major market effects.

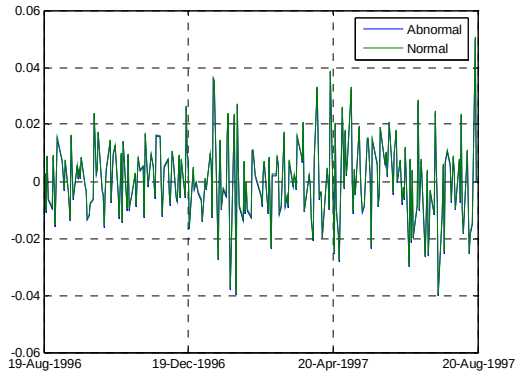


Figure 19: Comparison Open-to-Close returns for both normal and abnormal returns.

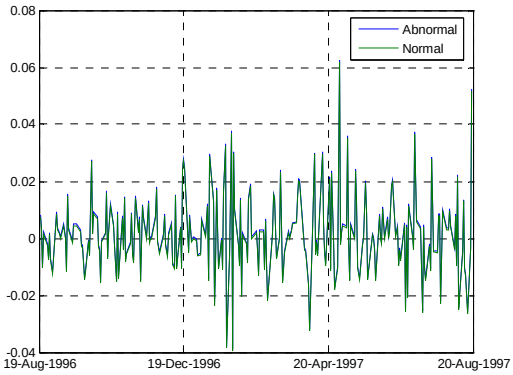


Figure 20: Comparison Open-to-Open returns for both normal and abnormal returns

In Table 3 a basic set of statistics are shown of the collection of returns being used for the experiment. The maximum, minimum, average and standard deviation are shown. The returns for Open-to-Open are on average higher than the Open-to-Close returns. Contrary to the visual inspection the abnormal returns are on average higher than the normal returns, at least for Open-to-Open returns.

	Maximum	Minimum	Average	Standard Deviation
Unadjusted Returns (Open-to-Close)	0,050684237	-0,039720369	0,000465430	0,013684734
Unadjusted Returns (Open-to-Open)	0,061904762	-0,039380616	0,001260918	0,013548989
Abnormal Returns (Open-to-Close)	0,050165456	-0,040223543	-0,000047380	0,013682548
Abnormal Returns (Open-to-Open)	0,062531653	-0,038759880	0,001882396	0,013551116

Table 3: Basic statistics of the returns data set.

To create the labels a cut-off for the division of the classes have to be chosen. This was already mentioned in section 3.2 where the parameter called gamma (γ) was mentioned. A logical first choice would be to set γ equal to 0, meaning that all the days with positive returns will be classified as up and all the days with negative returns will be classified as down. Another way of dividing the data into classes was to set the γ equal to the average plus and minus the standard deviation. This was done to specifically select only those days that have extraordinary returns compared to the rest of the returns used. To further simplify the problem instead of three classes (Up, Down and Neutral) only two classes (Up and Down) will be used. In Table 4 the number of days labelled are given with γ set to 0 and γ set to the average plus and minus the standard deviation as described above. For the actual values of γ set to the average plus and minus the standard deviation, see Table 5.

	$\gamma = 0$			γ = Average ± Standard Deviation		
	Down	**Up**	**Total**	**Down**	**Up**	**Total**
Normal Returns (Open-to-Close)	120	126	246	30	37	67
Normal Returns (Open-to-Open)	111	125	236	29	32	61
Abnormal Returns (Open-to-Close)	126	126	252	30	37	67
Abnormal Returns (Open-to-Open)	110	142	252	29	32	61

Table 4: Number of days labelled for each class and the total number of days.

	γ Values	
	Average - Standard Deviation	**Average + Standard Deviation**
Normal Returns (Open-to-Close)	-0,013219304	0,014150163
Normal Returns (Open-to-Open)	-0,012288071	0,014809907
Abnormal Returns (Open-to-Close)	-0,013729929	0,013635167
Abnormal Returns (Open-to-Open)	-0,011668720	0,015433512

Table 5: Actual values of γ set to average ± standard deviation.

4.3.2 Labelling by Trading Volume

The purpose of labelling by trading volume is to examine if certain news messages affect the trading behaviour of investors. The data will then be divided into two classes, high volume and low volume. The division is determined by either above or below the average trading volume or above/below the average trading volume plus/minus the standard deviation over the whole period. As in Table 3 basic statistics are shown of the acquired dataset in Table 6. In Table 7 the number of days is shown for both types of divisions.

	Maximum	Minimum	Average	Standard Deviation
Volume	6113000	533400	2249119	993578

Table 6: Basic statistics of the trading volume dataset.

	Average			Average ± Standard Deviation		
	Low	**High**	**Total**	**Low**	**High**	**Total**
Volume	143	109	252	35	47	82

Table 7: Number of days labelled by trading volume.

4.3.3 News Message Selection

As mentioned earlier the Reuters Corpus Volume 1 consists of over 800.000 news announcements in NewsML format. Due to the huge size of the corpus a selection was made to avoid computational problems and to make it easier to perform the experiment proposed in this thesis.

Reuters has coded the messages with Topic, Region and Industry codes. As the choice was made to use data from Exxon Mobil Corp., which is mainly dealing with Gas and Oil, it was obvious to make an initial selection of the news by only selecting from the Gas and Oil industry. In Appendix H a list of the industry codes are given that is used by Reuters. The codes used for selection, which were deemed most relevant, were made bold. After this selection procedure 30.613 news announcements remain covering 365 days.

The next step in filtering was to reduce the number of days to the number of trading days. This meant a reduction from 365 days to 254 days. Eventually we ended up using only 252 days, because one day was lost due to calculating returns and another day was lost due to a stock split. In the 252 days there were still 28.950 messages left, which was deemed substantial enough for experimenting. Due to this selection procedure a major assumption was made: news on weekends and holidays were disregarded, whereas the returns were not. Another major assumption which may prove problematic was that all the news messages on a certain day were assumed to have an effect on the returns of that day.

4.3.4 Fingerprinting

After having reduced the amount of news announcements from roughly 800,000 to about 29,000, the next step is to begin the fingerprinting process. As the software used is a commercial package, details on how it works can't be disclosed. To be able to use Collexis® to start fingerprinting, a lexicon/thesaurus is necessary.

A lexicon is a list containing concepts, whereas a thesaurus includes minor details on the relationship between concepts. Whatever is used the concepts that are contained within the lexicon or thesaurus are considered by the Collexis® Engine as the known reality or known world. Anything encountered within the list exists and that which isn't in the list does not exist at least not for the fingerprinting engine. To create this list of concepts the Term Extraction System mentioned by Van Eck (2005)[37] was used. This system extracts not only single word concepts, but also concepts of 2 words, 3 words or more. After running the application on the reduced corpus a number of terms were extracted. In Table 8 the number of terms extracted is shown by word length. To reduce the huge amount of terms identified, a choice was made to only select those terms that appear more than 100 times within the corpus. This is yet another arbitrary choice to reduce the computational complexity as mentioned earlier as one of the considerations of working with knowledge representations. After this selection process only 1,554 concepts were selected to be included in the lexicon.

Word length	Number of terms	Cut-off 100
1	22007	1109
2	91904	340
3	91065	84
4	62527	13
5	31284	6
6	12966	1
7	4921	1
8	1585	0
9	516	0
10	197	0
11	71	0
12	18	0
13	3	0
14	1	0
Total	**319065**	**1554**

Table 8: Number of terms extracted and the number of terms selected by cutting off at a frequency of 100.

Next to previous lexicon another was made by choosing concepts by a lower cut-off. In Table 9 you can see the number of terms selected by selecting only 1-word, 2-word and 3-word terms appearing more than 10 times in the corpus. As longer length terms were seen to be composed of lesser length terms, only 1-word, 2-word and 3-word terms were chosen. This procedure resulted in 197,231 concepts in the lexicon.

Word length	Number of terms	Cut-off 10
1	22007	14262
2	91904	91904
3	91065	91065
Total	**204976**	**197231**

Table 9: Number of terms extracted and the number of terms selected by cutting off at a frequency of 10.

Having created these lexicons the fingerprinting process is the next step. Indexing, as fingerprinting is called within the Collexis® software, was performed on the reduced corpus by using the two lexicons and the freetext thesaurus, which is an attempt of the engine to identify the concepts by itself. Next to a lexicon or thesaurus, a list of stop words need to be supplied, these are words that are not going to be disregarded by the Collexis® engine. This list can be found in Appendix I.

Lexicon/Thesaurus	Sources	Concepts	Average Vector Length
Lexicon 1	28,950	1,508	42
Lexicon 2	28,950	176,662	104
Freetext	28,950	73,546	106

Table 10: Statistics of Collexis®.

In Table 10 some statistics are shown after the indexing process. There were in total 28,950 news messages that were imported into the Collexis® software for fingerprinting. Using the first lexicon the engine identified 1,508 concepts, with the second lexicon 176,662 and finally with the freetext option 73,546. To focus this research it was chosen to only look at the fingerprints created by the first and smaller lexicon.

4.4 Dataset

Having now created the labels and the fingerprints, the fingerprints will be labelled with the various labels mentioned earlier by looking at the date of that fingerprint to create the dataset for modelling. There are in total 10 labels, 8 created by looking at returns and 2 by looking at trading volume. This means that there will be 10 datasets. In Table 11 the descriptions are given of the 10 created datasets as described earlier. For creating the model 10% of the dataset was randomly selected and taken out for out-of-sample testing, leaving 90% for training the model.

Dataset	Based on	Time Horizon	γ
1	Returns	Open-to-Close	0
2	Returns	Open-to-Close	**Average ± Standard Deviation**
3	Returns	Open-to-Open	0
4	Returns	Open-to-Open	**Average ± Standard Deviation**
5	Abnormal Returns	Open-to-Close	0
6	Abnormal Returns	Open-to-Close	**Average ± Standard Deviation**
7	Abnormal Returns	Open-to-Open	0
8	Abnormal Returns	Open-to-Open	**Average ± Standard Deviation**
9	Trading Volume	-	0
10	Trading Volume	-	**Average ± Standard Deviation**

Table 11: Description of the created datasets.

It seems prudent to take a closer look at the labels of the datasets. This produced Table 12, which contains the number of similar labels between the various return-based datasets. In Table 4 it is shown how many days are available for labelling. As mentioned earlier every day, defined either as open-to-close or open-to-open, gets a label up (+1) or down (-1). Table 12 then shows how many days between the datasets got the same label. Take, for example, the intersection between Dataset 2 and 5, which gives 45. This means that 45 days of both datasets have had the same labels assigned.

Dataset	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	-	73	**111**	28	**127**	*73*	120	28
2	73	-	34	**155**	45	***252***	31	155
3	**111**	36	-	77	**142**	36	**235**	77
4	28	**155**	77	-	57	**155**	61	***252***
5	**127**	45	**142**	57	-	45	**134**	57
6	*73*	***252***	36	**155**	45	-	31	**155**
7	120	31	**235**	61	**134**	31	-	61
8	28	155	77	***252***	57	**155**	61	-

Table 12: Number of similar labels between the datasets created by returns.

The numbers in bold were marked as those interesting to check, as these datasets differed on only one variable. The numbers both bold and italic mean that the two datasets are similar, which means that all the days for both datasets have exactly the same labels. It has to be noted that Dataset 2 has the same labels as Dataset 6 and Dataset 4 the same as Dataset 8. This fact would seem to indicate that between returns and abnormal returns with a γ equal to average returns ± standard deviation, it doesn't matter if you use normal returns or abnormal returns. This could be consistent with what was mentioned earlier in section 4.3.1, when we noted the low R^2.

Dataset	**Train**	**Test**
1	25470	2842
2	6840	763
3	24489	2729
4	6633	742
5	25038	2785
6	6840	763
7	26046	2904
8	6633	742
9	26046	2904
10	8190	907

Table 13: Total number of instances for each dataset.

Also mentioned earlier, the data retrieved was of a huge quantity. One of the major concerns with this was the time necessary to process all the data. Therefore, the total number of instances is shown in Table 13 for each dataset, divided into the number of instances for training and testing. Every instance consists of a fingerprint, which is a representation of a news message, and a label, which represents the stock movement. As can be seen, when a γ of 0 is used there are almost 29000 instances, whereas a γ equal to average returns ± standard deviation has around 8000 instances.

In the next chapter the choices pertaining to the SVM algorithm will be discussed. The results of the predictive power of the trained model by SVM will also be shown. Preliminary conclusions will also be given.

5 RESULTS

"Results! Why, man, I have gotten a lot of results. I know several thousand things that won't work."

--Thomas A. Edison (1847 - 1931)

"The sciences do not try to explain, they hardly even try to interpret, they mainly make models. By a model is meant a mathematical construct which, with the addition of certain verbal interpretations, describes observed phenomena. The justification of such a mathematical construct is solely and precisely that it is expected to work.."

--Johann Von Neumann (1903 - 1957)

5.1 Introduction

As mentioned at the end of the previous chapter, a description of the model used for training and testing will be given. Not just the choices are given, but also the process of how the model used was selected will be given. This chapter will show the classification rate or accuracy of the models for the various datasets on both the training as well as the test datasets.

5.2 Kernel Selection

To perform the SVM algorithm as described in chapter 3 there are certain choices that need to be made. This algorithm comes in many forms. For experimentation purposes it was chosen to use MATLAB Toolbox OSU SVM Classifier[48] written by Junshui Ma, Yi Zhao, and Stanley Ahalt, which core is based on LIBSVM Algorithm[49] by Chih-Chung Chang and Chih-Jen Lin.

[48] http://www.ece.osu.edu/~maj/osu_svm/
[49] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

One of the first choices that had to be made was the kernel function. There are a number of kernels that can be used in SVM models. These include linear, polynomial and radial basis function (RBF). Another choice is the form of the error function. SVM models for Classification can be classified into two groups, Type 1 (also know as *C*-SVM) and Type 2 (also known as *v*-SVM).

Kernel	**Type**	**Dataset 1**		**Dataset 2**	
		Train	**Test**	**Train**	**Test**
Linear	C-SVM	61,43%	52,50%	73,54%	55,83%
	v-SVM	53,05%	50,74%	72,87%	56,62%
Polynomial	C-SVM	98,35%	56,26%	99,17%	61,47%
	v-SVM	94,96%	56,65%	95,98%	61,99%
RBF	C-SVM	96,69%	54,64%	98,08%	59,24%
	v-SVM	97,80%	55,56%	98,58%	60,16%

Table 14: Classification Rate using various kernels with both *C*-SVM and *v*-SVM.

First to find the appropriate kernels, the three different kernel types were run on 2 datasets with default settings using both error functions. The results to this are shown in Table 14 by giving the classification rate. This rate is defined as the percentage of instances that were correctly classified. It can be seen then, that both Polynomial and RBF kernels perform much better than linear kernels, indicating a linear plane to be insufficient for separation. The results of the two types of SVM models for the Polynomial and RBF kernels show the *v*-SVM model to give better results. As the polynomial kernel gives better classification rates on the test set, it has been chosen to use the polynomial function for the kernel. Ultimately, we are supposed to be able to accurately classify new and unseen instances.

As the RBF kernel is by far the most popular choice of kernel types used in SVM and utilises a lot less computational time than a polynomial function, it was chosen to use this kernel as well. The much shorter time may be preferred as the results are similar to those of a polynomial kernel.

5.3 Determining parameters for Polynomial kernel

	Dataset 1		Dataset 2	
Degree	**Train**	**Test**	**Train**	**Test**
1	52,76%	51,86%	72,28%	57,27%
2	94,97%	55,74%	95,69%	60,55%
3	94,96%	56,65%	95,98%	61,99%
4	94,84%	55,28%	95,75%	61,34%
5	94,92%	54,82%	94,59%	60,42%

Table 15: SVM run on Dataset 1 and Dataset 2 using Polynomial kernel function with different *Degree*.

The OSU SVM Toolbox requires one parameter to train this classifier, *degree*. *Degree* determines the degree of the polynomial function. In Table 15 the results for using different *Degree* to train and test Dataset 1 and 2 are shown. The default setting for the toolbox in MATLAB was 3 for *Degree*. Seeing these results it has been chosen to stick with the default. It should also be noted that *Degree* equalling 1 has similar results as the linear kernel function.

5.4 Determining parameters for RBF kernel

The OSU SVM Toolbox used requires two parameters to train this classifier, *Gamma* and *C*. *Gamma* is a RBF parameter and *C* is the cost of the constrain violation. Table 16 shows the results for using different *Gamma* and *C* to train and test Dataset 2. The default setting for the toolbox in MATLAB was both 1 for *Gamma* and *C*. Seeing these results it has been chosen to stick with the default. A different *C* does not seem to have any effect on the classification rate.

		Dataset 2	
Gamma	**C**	**Train**	**Test**
1	1	98,58%	60,16%
1	10	98,58%	60,16%
1	100	98,58%	60,16%
1	1000	98,58%	60,16%
1	10000	98,58%	60,16%
1	100000	98,58%	60,16%
0,1	1	92,35%	59,11%
0,1	10	92,35%	59,11%
0,1	100	92,35%	59,11%
0,1	1000	92,35%	59,11%
0,1	10000	92,35%	59,11%
0,1	100000	92,35%	59,11%
0,01	1	66,05%	55,44%
0,01	10	66,05%	55,44%
0,01	100	66,05%	55,44%
0,01	1000	66,05%	55,44%
0,01	10000	66,05%	55,44%
0,01	100000	66,05%	55,44%
0,001	1	59,66%	55,31%
0,001	10	59,66%	55,31%
0,001	100	59,66%	55,31%
0,001	1000	59,66%	55,31%
0,001	10000	59,66%	55,31%
0,001	100000	59,66%	55,31%
0,0001	1	58,20%	51,11%
0,0001	10	58,20%	51,11%
0,0001	100	58,20%	51,11%
0,0001	1000	58,20%	51,11%
0,0001	10000	58,20%	51,11%
0,0001	100000	58,20%	51,11%
0,00001	1	45,13%	44,95%
0,00001	10	45,13%	44,95%
0,00001	100	45,13%	44,95%
0,00001	1000	45,13%	44,95%
0,00001	10000	45,13%	44,95%
0,00001	100000	45,13%	44,95%

Table 16: SVM run on Dataset 2 using Rbf kernel function with different *Gamma* and *C*.

5.5 Naïve Classifier

After having determined the kernel and the parameters to use, it was suggested to use a naïve classifying system as a benchmark for comparison. This naïve system would classify the test set by the most frequent class of the training set.

To do this type of classification the distribution of the classes must first be known for the various datasets. In Table 17 the distribution is shown of the instances over the two classes and the total number of instances for training and testing for each dataset.

	Train			Test		
Dataset	**1**	**-1**	**Total**	**1**	**-1**	**Total**
1	12852	12618	25470	1436	1406	2842
2	3753	3087	6840	420	343	763
3	13167	11322	24489	1469	1260	2729
4	3600	3033	6633	402	340	742
5	7533	17505	25038	840	1945	2785
6	3753	3087	6840	420	343	763
7	14814	11232	26046	1654	1250	2904
8	3600	3033	6633	402	340	742
9	11358	14688	26046	1268	1636	2904
10	4896	3294	8190	545	362	907

Table 17: Number of instances per class and the total number for training and test set.

In Table 18 the results are given for a naïve classification system. As mentioned earlier, it predicts the class which appeared most frequent in the training set. The results indicate that the datasets are not balanced, which means that the distribution of the instances over the classes are not equal. If the distribution of the instances over the two classes were equal, there would be a 50% chance for either class. This fact can also be seen in Table 17.

	Naïve Classifier		
Dataset	**Train**	**Test**	**Prediction**
1	50,46%	50,53%	+1
2	54,87%	55,05%	+1
3	53,77%	53,83%	+1
4	54,27%	54,18%	+1
5	69,91%	69,84%	-1
6	54,87%	55,05%	+1
7	56,88%	56,96%	+1
8	54,27%	54,18%	+1
9	56,39%	56,34%	-1
10	59,78%	60,09%	+1

Table 18: Results of a naïve classifying system.

5.6 Results for polynomial kernel classification

90% of each dataset was used for training and 10% was left out for testing. After a long time of solving out-of-memory problems and then finally waiting for the model to be trained, both the classification rate was determined on the training set as well as the test set. The results are given in Table 19.

Dataset	Classification Rate	
	Train	Test
1	94,96%	56,65%
2	95,98%	61,99%
3	94,81%	56,65%
4	96,26%	59,16%
5	85,87%	70,16%
6	96,08%	61,07%
7	93,88%	56,71%
8	96,43%	63,48%
9	93,85%	69,35%
10	94,85%	74,31%

Table 19: Classification Rate using polynomial kernel on the various datasets.

As can be seen the trained classifier performs well on the training set. The test rate is however a lot lower, but still much better than guessing. It also performs better than the naïve classifier by almost 5%. Something interesting to note would be the results for Dataset 5 and 7. The classification rate for Dataset 5, which is the dataset with return labels created by abnormal returns from open-to-close with γ equal to 0, seems to only do 2% better than the naïve classifier. The rate on Dataset 7, which is the dataset with return labels created by abnormal returns from open-to-open with γ equal to 0, however, is doing even worse than the naïve classifier.

Another point to be noted is that the classification performed on the trading volume prediction is much better than on returns. This difference of classification can be almost 5% higher than on returns. This would seem to suggest that predicting trading behaviour would be easier than predicting stock price movement.

5.7 Results for Rbf kernel classification

The polynomial kernel had the highest classification rate on the test set. The Rbf kernel, however, is more popular. This popularity is mostly due to the computational time of this algorithm. The results for using this kernel are shown in Table 20.

Dataset	Classification Rate	
	Train	Test
1	97,80%	55,56%
2	98,58%	60,16%
3	97,82%	56,87%
4	98,82%	56,06%
5	98,11%	71,17%
6	98,71%	58,72%
7	97,80%	58,78%
8	98,69%	60,78%
9	98,52%	66,25%
10	99,24%	68,91%

Table 20: Classification Rate using Rbf kernel on the various datasets.

As can be seen this trained classifier performs well on the training set, almost fitting perfectly. The test rate is however a lower, but still better than guessing. Overall the performance is about 2-5% higher than the naïve classifier. Compared to the results of the polynomial kernel, this kernel always performs better than a naïve classifier.

The same note can be made about using the trading volume for prediction as with the polynomial kernel. SVM performs much better on trading volume than on returns prediction.

5.8 Comparison

To compare all the results from the previous sections for the naïve classifier, SVM classifiers using polynomial kernel and Rbf kernel, the results are given in Table 21.

Dataset	Train			Test		
	Rbf	Poly	Naïve	Rbf	Poly	Naïve
1	97,80%	94,96%	50,46%	55,56%	56,65%	50,53%
2	98,58%	95,98%	54,87%	60,16%	61,99%	55,05%
3	97,82%	94,81%	53,77%	56,87%	56,65%	53,83%
4	98,82%	96,26%	54,27%	56,06%	59,16%	54,18%
5	98,11%	85,87%	69,91%	71,17%	70,16%	69,84%
6	98,71%	96,08%	54,87%	58,72%	61,07%	55,05%
7	97,80%	93,88%	56,88%	58,78%	56,71%	56,96%
8	98,69%	96,43%	54,27%	60,78%	63,48%	54,18%
9	98,52%	93,85%	56,39%	66,25%	69,35%	56,34%
10	99,24%	94,85%	59,78%	68,91%	74,31%	60,09%

Table 21: Comparison of the results for SVM and Naïve Classifier.

Looking at the results in Table 21, this modelling gives results that look very promising. Overall this experiment has shown that interpreting news does add value or predictive power to market movement prediction.

5.9 Confidence Interval[50]

Dataset	Rbf		Poly		Naïve	
1	97,62%	97,98%	94,69%	95,23%	49,85%	51,07%
2	98,30%	98,86%	95,51%	96,45%	53,69%	56,05%
3	97,64%	98,00%	94,53%	95,09%	53,14%	54,39%
4	98,56%	99,08%	95,80%	96,72%	53,08%	55,47%
5	97,94%	98,28%	85,44%	86,30%	69,35%	70,48%
6	98,44%	98,98%	95,62%	96,54%	53,69%	56,05%
7	97,62%	97,98%	93,59%	94,17%	56,27%	57,48%
8	98,42%	98,96%	95,98%	96,88%	53,08%	55,47%
9	98,37%	98,67%	93,56%	94,14%	55,79%	56,99%
10	99,05%	99,43%	94,37%	95,33%	58,72%	60,84%

Table 22: Confidence Interval for the training set with 95% probability.

To check how good the estimated classification rate (*Classification rate*$_e$) is compared to the actual classification rate, we can calculate confidence intervals. As the number of instances is much larger than 30, it can be assumed as suggested in Mitchell (1997)[50] that the probability distribution is normally distributed.

[50] Mitchell, T.M., 1997, "*Machine Learning,*" Singapore, McGraw Hill International Editions, Computer Science Series, pp. 128-153.

In Table 22 and Table 23 is shown that with approximately 95% probability, the true classification rate lies for the training and test set respectively in the interval

$$Classification\ rate_e \pm 1{,}96 * \sqrt{(\ Classification\ rate_e * (1 - Classification\ rate_e) / n\ )}$$

where $Classification\ rate_e$ is the estimated classification rate and n is the number of instances.

Dataset	Rbf		Poly		Naïve	
1	53,73%	57,39%	54,83%	58,47%	48,69%	52,37%
2	56,69%	63,63%	58,55%	65,43%	51,52%	58,58%
3	55,01%	58,73%	54,79%	58,51%	51,96%	55,70%
4	52,49%	59,63%	55,62%	62,70%	50,59%	57,76%
5	69,49%	72,85%	68,46%	71,86%	68,13%	71,54%
6	55,23%	62,21%	57,61%	64,53%	51,52%	58,58%
7	56,99%	60,57%	54,91%	58,51%	55,16%	58,76%
8	57,27%	64,29%	60,02%	66,94%	50,59%	57,76%
9	64,53%	67,97%	67,67%	71,03%	54,53%	58,14%
10	65,90%	71,92%	71,47%	77,15%	56,90%	63,28%

Table 23: Confidence Interval for the test set with 95% probability

For the training set the results are the same as previously stated, which means that the model is indicating probable overfitting. Looking at the test set results, however, show that the actual classification rate of Dataset 5 is similar for all three methods of classification. This would indicate that the model created by Dataset 5 predicts very badly. When comparing to the results of Dataset 1, we could argue that normal returns should be used above abnormal returns, but as mentioned earlier, seeing the low R^2, the abnormal returns were already considered unreliable.

5.10 Discussion

In the previous sections of this chapter various results have been given and some peculiarities have already been highlighted. As mentioned earlier there appears to be a difference between the results of Datasets 1 to 8 and Datasets 9 and 10. The latter has overall better results. This suggests that predicting trading behaviour by using news is more successful than predicting returns by news.

Another important point that needs highlighting is that using a γ not equal to 0 gives overall better results. This can be explained by the fact that by doing this only the really abnormal movements are used. Intuitively, this would mean that only those days that have special price movements are retained and matched with news.

The proposition if open-to-open or open-to-close returns were better shows that open-to-close returns seem to have more predictive power. This is, however, only for using normal returns. But as abnormal returns were deemed unreliable, this would deserve a closer examination.

The bad results attained by using Dataset 5 and 7 should also be examined closer. Initially, it would indicate that γ equal to 0 should not be used for either open-to-open or open-to-close returns.

Finally, it should be noted that except for Datasets 3, 5 and 7, the polynomial kernel performs better than the Rbf kernel. The results of Datasets 5 and 7 are dubious at best. It can therefore be concluded that polynomial kernel is truly better. A side note which has to be mentioned is that the modelling time using the polynomial kernel, however, was almost 4 to 5 times more. In appendix J confusion matrices are given for both types of kernels for all 10 datasets. It shows that overall polynomial kernel has a higher percentage of correctly classifying both classes, whereby Rbf kernel often classifies only one of the two classes very well.

6 CONCLUSION

"Statistics: The only science that enables different experts using the same figures to draw different conclusions."

--Evan Esar (1899 - 1995)

"A conclusion is the place where you got tired of thinking."

--Harold Fricklestein

6.1 Introduction

After showing the results in the previous chapter, this chapter will take another look at the total experiment by first recapping. Conclusions will then follow by attempting to answer the questions opted in the first chapter. This thesis will further be concluded by giving suggestions for future research.

6.2 Recap

In this thesis an attempt was made to find the effect or influence of news on the market. The question was also asked if by being able to interpret news messages a better prediction of market movements could be made. The problem was conceptually depicted in Figure 1. An extensive literature survey was done and a possible solution was found by transforming the problem into a simpler classification problem, whereby representations were made for both news and market and then matched. This was shown in Figure 4. The purpose for finding this effect of news was to be able to make predictions of the market, hoping this ability to predict could help achieve higher returns.

The prediction process was then given in Figure 5. First the market data was transformed into class labels and then news into fingerprints. Fingerprints were defined as lists of concepts accompanied by a score. The fingerprints were then labelled giving every fingerprint a class label (Up / Down), indicating the market movement.

By this procedure a total of 10 datasets were created, whereby after examination 2 datasets were duplicates. The datasets were different due to the different labels assigned to them. These datasets were then split into training and test set with a 9 to 1 ratio. The training set was then used to train a SVM Classifier, which was then tested on both the training and test set for its prediction accuracy. (See Figure 21 for a diagram of the process of experimentation in this thesis.

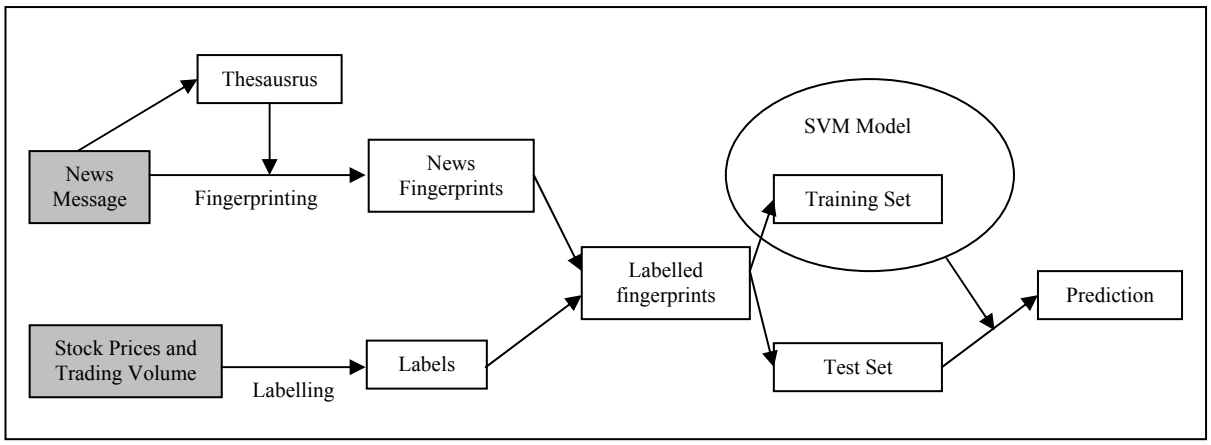


Figure 21: Process of Experimentation in this thesis.

6.3 Conclusions

In the previous chapter the results were given of the classifiers by the classification rate or also called accuracy. The results were deemed promising indicating that using news could boost the power to predict. This would mean that news truly had an effect on market movements, which was already established in earlier works.

Why the results are given such a positive light is because of the data that was used in this thesis. Daily returns or trading volumes were matched with messages of a single to day, which amounted to almost 30.000 messages spread over 252 trading days. Keeping this in

mind, when using such low frequency data and still being able attain such promising results would indicate that further increase of granularity or using higher frequency data would surely give even better results.

The answer to the other question asked about the ability to interpret news messages for better market predictions is also deemed to be positive seeing that the accuracy is mostly above that of a naïve classifier. This would indicate that using this system of prediction a better than random result could be achieved. It shows that using fingerprints boosts the predictive power.

In this thesis a very good and accurate system for prediction was not found. However, it did add to the existing literature by defining accurately the problem to be solved and using different algorithms in an already existing methodology. The use of Collexis® Fingerprints in this area of research combining both information retrieval technology with economics in general and finance in particular is deemed as a new approach. The choice of using SVM on the other hand for classification was based on literature studied as mentioned in chapter 2. The many choices made in this thesis can also serve as a good starting point for future research, which will be elaborated in the following section.

6.4 Further Research

Suggestions for further research are many. In this process of finding a representation for both the market and news, matching both representations, training the model and then testing it, a lot of choices were made. Some of these choices were arbitrary, but others were made based on research and experimentation. For every step in the process of this thesis choices were made and these choices would deem closer examination in future work.

But even before beginning the process choices were already made. One of the major choices made, was in the data selection process. The news articles retrieved were without timestamp, which meant that all the announcements were assigned to 1 day. This would ultimately mean that over 1.000 articles got a single label as it belonged to a certain day. So the first suggestion would be that if articles of greater granularity, which means more

separation or higher frequency, were used better results may be attained. As mentioned within this work, intraday prices and news announcements would have been preferred, but that kind of data was either not available or only retrievable at a cost. Even without tuning any other choices in this experiment the choice of the granularity of the news articles is believed to be able to significantly improve the results.

Getting back to the procedure, the first step in the procedure then was the choice of representations for the market and news. It was chosen to use stocks as an instance of the market. Other choices like currency, options, forwards, futures should also be examined. It was then chosen to use prices and trading volume for labelling. Returns were then calculated from the prices. Moving averages, volatility or variance are suggestions for further research. The choice of using abnormal returns instead of normal returns suggested in literature didn't payoff in this experiment due to the choice of using S&P 500 as a benchmark of the market. Another benchmark could and should be used to see if as literature suggests abnormal returns reflect better the company specific price movements. When creating the labels by the returns and trading volume, many other choices were made to contain the size of this experiment. The choices for γ were limited to 2, whereas many other γ's could have been chosen. It could also be suggested to use more than 2 classes as was initially proposed in this thesis.

This thesis was born out of the suggestion of using Collexis® Fingerprints for the representation of news. Therefore fingerprints were chosen as are representation of news. The algorithms for the creation of these fingerprints were set to default for this experiment, but it would be interesting to examine the results by using other algorithms. Another factor which is of importance in the creation of fingerprints is the lexicon that was created. The lexicon created could be enriched to make a thesaurus by including basic relationships between the found concepts. If time and experts in the field of oil and gas were available, an even more extensive and accurate thesaurus could be created, which give similarity and specificity values more weight allowing the use of the different other algorithms. This would have given the fingerprints an even higher accuracy of representing the news articles. In chapter 2 many other representations were shown which were used in literature and seeing the growing complexity of structures in technology, new and fuller representations may be found in the future as technology advances.

The choice of the algorithm for training and testing was also an important factor in this research experiment. It was chosen to use SVM, but other algorithms like neural networks, decision trees or *k*-nearest neighbour could also have been used to solve this classification problem. Even after choosing SVM the right parameters could be further examined. In this study a basic search was done to choose the kernel and the parameters for the kernel, but a more extensive search could enhance the results. The SVM should be further analysed to see if a relationship could be found between the fingerprints and the label given. This may result in highlighting certain concepts which regularly have a certain effect on the market.

Finally, making a classification problem to solve the problem opted in this thesis was the first choice and was found often in literature. But this does not mean that there may not be any other ways to examine or research the original problem. Looking at other sciences and their methodologies may or may not lead to even more creative solutions to this problem.

Bibliography

Anderson, J. R., and Bower, G. H., 1973, *"Human Associative Memory*," New York, John Wiley and Sons, pp. 9.

Baestaens, D.J.E., and Van den Bergh, W.M., 1996, "*Public information Effects on the DEM/USD swap rate: An intraday analysis in operational time*," Rotterdam Institute for Business Economic Studies, R 9602/F, Erasmus University Rotterdam.

Blasco, N., Corredor, P., Del Rio, C., Santamaria, R., 2005, "Bad news and Dow Jones make the Spanish stocks go round," *European Journal of Operational Research* 163, pp. 253-75.

Brealey, R.A. and Myers, S.C., 2003, "*Principles of Corporate Finance,*" 7th ed., McGraw-Hill Higher Education, pp. 344-75.

Bunningen, A.H. van, 2000, "*Augmented Trading*," Master Thesis, University of Twente, Enschede, The Netherlands.

Cho, Y-H., and Engle, R.F., 2000, "*Time-Varying Betas and Asymmetric Effects of News: Empirical Analysis of Blue Chip Stocks*."

Davis, R., Shrobe, H., and Szolovits, P., 1993, "What is a Knowledge Representation?" *AI Magazine*, 14(1):17-33.

Eck, N.J. van, 2005, "*Towards Automatic Knowledge Discovery from Scientific Literature, Computer Based Tools for Supporting Scientific Research,*" Master Thesis, Erasmus University Rotterdam, The Netherlands.

Fleming, M.J., and Remolona, E.M., 1999, "Price Formation and Liquidity in the U.S. Treasury Market: The Response to Public Information," *Journal of Finance*, Vol. 54, No. 5, pp. 1901-15.

Funke, N., and Matsuda, A., 2002, "Macroeconomic News and Stock Returns in the United States and Germany" IMF Working Paper WP/02/239, (IMF Institute).

Hariharan, G., 2004, "*News Mining Agent for Automated Stock Trading*," The University of Texas at Austin, USA.

Hamburger, Y., 2004, "*The Exceptional Event*," Erasmus University Rotterdam, The Netherlands.

Hardouvelis, G.A., 1986, "Macroeconomic Information and Stock Prices," First Boston Working Paper Series FB-86-13, (New York: Columbia University).

Harmelen, F. van, and Fensel, D., 1999, "*Practical Knowledge representation for the Web*," IJCAI'99 Workshop on Intelligent Information Integration.

Hastie, T., Tibshirani, R., and Friedman, J., 2001, *The Elements of Statistical Learning; Data Mining, Inference, and Prediction,* Springer-Verlag, New York.

Iturres, A.S., 2001, "Market Prediction Technique," *New Trading Ideas Internet Journal*, Publication 01-07.

Jacobs, P.S., and Rau, L.F., 1990 , "SCISOR: Extracting Information from On-line News. An Object-Oriented Relational Database," *Communications of the ACM*, Vol. 33, No. 11, pp 87-97.

Kaminsky, G.L., and Schmukler, S.L., 1999, "What triggers market jitters: A Chronicle of the Asian Crisis," *International Finance Discussion Papers*, Number 634.

Kendall, M.G., 1953, "The Analysis of Economic Time Series, Part I. Prices," *Journal of the Royal Statistical Society* 96, pp. 11-25.

Li, L, and Hu, Z.F., 1998, "Responses of the Stock Market to Macroeconomic Announcements Across Economic States," IMF Working Paper 98/79 (Washington International Monetary Fund).

Mandelbrot, B.B., and Hudson, R.L., 2004, *The (Mis)Behaviour of Markets, A Fractal View of Risk, Ruin and Reward*, Basic Books, pp. 88-94.

Mitchell, M.L., and Mulherin, J. H., 1994, "The Impact of Public Information on the Stock Market," *The Journal of Finance*, Vol. 49, No. 3, Papers and Proceedings Fifty-Fourth Annual Meeting of the American Finance Association, Boston, Massachusetts, January 3-5, 1994, 923-50.

Mitchell, T.M., 1997, "*Machine Learning,*" Singapore, McGraw Hill International Editions, Computer Science Series, pp. 128-153.

Mueller, E.T., 2000, "*Making news understandable to computers*," Signiform, Washington, D.C.

Neches, R., Fikes, R., Finin, T., Gruber, T., Patil, R., Senator, T., and Swartout, W.R., 1991, "Enabling Technology For Knowledge Sharing," *AI Magazine*, Volume 12, No. 3.

Patell, J.M., and Wolfson, M.A., 1982, "Good News, Bad News, and the Intraday Timing of Corporate Disclosures," *The Accounting Review*, Vol 57, No. 3 (Jul., 1982), pp. 509-27.

Quillian, M. R., 1968, "Semantic Memory," In Minsky, M. (Ed.) (1968), *Semantic Information Processing,* Cambridge, Mass.: MIT Press, pp. 9, 216.

Sar, N.L. van der, 1997, "*Event-studies: Methodologische aspecten*," Vakgroep Financiering en Belegging, Erasmus Universiteit Rotterdam. [*Paper is in Dutch*].

Schölkopf, B., 1997, *Support vector learning*, GMD-Berichte No. 287, GMD-Forschungszentrum Informationstechnik.

Sowa, J.F., 2000, *Knowledge Representation: Logical, Philosophical, and Computational Foundations,* Brooks/Cole Publishing Co., Pacific Grove, CA.

Sun, Q., and Tong, W.H.S., 2000, "The Effect of United States Trade Deficit Announcements on the Stock Prices of United States and Japanese Automakers," *Journal of Financial Research*, Vol. 23, No. 1, pp 15-43.

Tan, F.H., 2005, "*Option Pricing, the GARCH-M Approach,"* Bachelor Thesis, Erasmus University Rotterdam, The Netherlands.

Vapnik, V., 1996, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York.

APPENDIX A: Semantic Relations

-Onym

Morphemes with the affix -onym (from the Greek for name) are designations for either a closed set of grammatical morphemes that refer to relationships between word pairs, such as synonym and antonym; or they may stand for classical compound nouns of an open type that refer to a particular subject, such as toponym, charactonym, etc. By analogy they may be freely created, sometimes for no other reason than to give an erudite impression of the user who expects his listeners to understand Greek, and it is in this way that words such as ornithonym or ichthyonym may be formed.

The usage of the word pairs is of great importance in grammar. Some morphemes ending in -onym may represent words that contain components, in the way house may contain window, roof, and door, or they may be words so contained in others, such as steering-wheel and engine in car. They may be generic words that stand for a class or group of equally-ranked items, such as tree for beech or elm, or belong within that class, such as lily or violet in flower. They may have the same or a similar meaning as a differently spelled word, such as sofa or couch, or they may stand in direct contrast to another, such as useful and useless. Some morphemes have the -nym form rather than the -onym form, such as ananym or hypernym, but that may be more for ease of pronunciation than for etymological reasons.

Most -onyms may have suffixes added to them and in this way form derivatives with the endings -onymy, -onymous, -onymic, etc., in new constructions. Others may reverse this process by removing suffixes in back-formations; especially if the new morphemes thus formed sound plausible enough to have been the root form in first place.

A list of -onym words

- acronym - an abbreviation formed from the initials of one or more words that is pronounceable like a normal word, such as NATO, sometimes in distinction to initialism
- allonym - an author's name of another person's, often a well-known person's name
- anacronym - portmanteau morpheme of anachronism + acronym that may be an acronym, abbreviation, or initialism but that is so well established that its origin is no longer remembered. As this seems a specious reason (after all any acronym may be looked up in a dictionary) it may be more applicable to abbreviations, etc., that are at risk of becoming out-of-date and being superseded by newer abbreviations, such as AD being replaced by CE (Common Era)
- ananym - a name written backward and used as pseudonym
- anonym - word defining anything created anonymously, or the person who has created it; an unknown author; a pseudonym
- anthroponym - a classical compound word denoting a human name
- antonym - one of the words of the word pair antonym and synonym that indicates the exact opposite meaning of another word, an antithesis, such as high and low
- aptronym - name appropriate to its owner's occupation or physical properties as in Goldsmith or Longman
- aristonym - one of the classical compound nouns of -onym words that denotes a name that is derived from a high rank or a title of nobility
- backronym - portmanteau morpheme of back + acronym that appears to fit an existing word but has really been created as an acronym, such as BASIC (Beginner's All-purpose Symbolic Instruction Code)
- caconym - one of the words of the word pair caconym and euonym; a word that is wrongly applied; a misnomer; the incorrect name for something, especially in the classification of plants, etc.

- capitonym - a compound word of capital + -onym of a word that changes its pronunciation and meaning when it is capitalized, and usually applies to capitalization due to proper names or eponyms, as in August - august, or Polish - polish
- charactonym - a compound word, the name of a fictional character that may be reflected in his personality traits, as in Shakespeare's Pistol or Bottom; or Titus Feuerfuchs in Johann Nestroy's Der Talisman, who attempts to hide his fiery-red hair with a black wig
- contronym - a word that may have opposite meanings in different contexts, such as to cleave - to stick to, adhere, and to cleave - to split
- cryptonym - a classical compound noun denoting a code name; a word or name used clandestinely to refer to another name or word
- demonym - one of the classical compound nouns, a name of persons/people that refers to the place they come from, such as the Assyrian, or the Briton
- eponym - a botanical, zoological or place name that derives from a real or legendary person; a name for a real of hypothetical person from whom a botanical, geographical or zoological name is derived; a person after whom a medical condition is named, or the condition so named
- ethnonym - a classical compound word denoting a name of an ethnic group
- euonym - one of the words of the word pair euonym - caconym; a word well suited to a person, place or thing so named; a pleasant name
- exonym - a name used by one group of people for another group, but who call themselves by a different name, such as the name Germans used by English-speakers for Deutsche, the name German-speakers use. Other examples are city names such as Cologne in English whose German equivalent is Köln
- heteronym - one of the words of the word pair homonym and heteronym; a word that is spelled in the same way as another but that has a different sound and meaning, such as bow of the ship and bow and arrow.
- holonym - one of the words of the word pair holonym and meronym; a word for the whole and of which other words are part, in the way house contains roof, door and window; or car comprises steering-wheel and engine
- homonym - one of the words of the word pair homonym and heteronym, or the word pair homonym and isonym; word that is pronounced and spelt the same way as another, but has a different meaning, such as bat, the mammal, and bat, the club
- hydronym - a classical compound word denoting a name of a body of water

- hypernym/hyperonym - one of the words of the word pair hypernym and hyponym; a generic type of word that stands for a class or group of equally-ranked items, such as tree for beech or elm, or house for chalet or bungalow. A hypernym is superordinate to a hyponym.
- hyponym - one of the words of the word pair hyponym and hypernym; an item that belongs to and is equally-ranked in a generic class or group, such as lily or violet in the class of flowers; or limousine or hatchback in the class of automobiles. A hyponym is subordinate to a hypernym
- isonym - often one of the words in the word pair homonym and isonym; word that is spelt the same as another word but sounds differently; or is of the same derivation as another and is therefore a cognate of that word
- meronym - one of the words of the word pair meronym and holonym; a word that names a part that belongs to and is therefore subordinate to a larger entity; a part - whole relationship, such as door or window in house, or engine or steering-wheel in car
- metonym - a word that substitutes a part for the whole it is associated with, such as crown for monarch. It is associated with its derivative, the figure of speech, metonymy
- paronym - a word that is related to another word and derives from the same root; a cognate word, as in dubious and doubtful
- patronym or patronymic - a classical compound word denoting a name adopted from the father's or ancestor's name, such as Johnson - John's Son; MacDonald - Son of Donald; O'Brien - Son of Brien; Ivanov - Son of Ivan, etc.
- pseudonym - a false and fictitious name adopted by an author; a pen name
- retronym - a replacement of an original simple noun by a modified noun, by having one or more components added to it, as in watch that existed on its own originally and then had a preceding analog added to it, in order to differentiate it from a digital watch
- synonym - one of the words of the word pair synonym and antonym; a word equivalent in meaning or nearly so to another word; a word that may be substituted for another word that has the same or a similar meaning, such as near and close
- tautonym - a binomial or scientific name in the taxonomy of animals in which the generic and specific names are the same as in Gorilla gorilla; a scientific name in which the specific name is repeated, as in Homo sapiens sapiens as distinct from Homo sapiens neanderthalensis; a noun component that is repeated, such as aye-aye or tom-tom; a personal name where both forename and surname are identical as in Francis Francis

- toponym - a classical compound noun that stands for a place or geographical name; the name of an area of the body, as distinguished from the name of an organ
- troponym - a verb that indicates more precisely the manner of doing something by its replacing a verb of a more generalized meaning, thus the verb to stroll indicates a more leisurely, casual manner of to walk.

APPENDIX B: Domain Knowledge

Dictionary

A dictionary is a list of words with their definitions, a list of characters with its glyph or a list of words with corresponding words in other languages. Many dictionaries also provide pronunciation information, word derivations, histories, or etymologies, illustrations, usage guidance, and examples in sentences.

Lexicon

A lexicon is a list of words together with additional word-specific information, i.e. a dictionary.

In linguistics, a lexicon has a slightly more specialized definition, as it includes the lexemes used to actualize words. Lexemes are formed according to morpho-syntactic rules and express sememes. In this sense, a lexicon organizes the mental vocabulary in a speaker's mind: First, it organizes the vocabulary of a language according to certain principles (for instance, all verbs of motion may be linked in a lexical network) and, second, it contains a generative device producing (new) simple and complex words according to certain lexical

rules. For example, the suffix '-able' can be added to transitive verbs only such that we get 'read-able' but not '*cry-able'.

Furthermore an individual lexical knowledge (or lexical concept) is a term used in academia to refer to an individual's vocabulary knowledge.

Ontology

In computer science, an ontology is the attempt to formulate an exhaustive and rigorous conceptual schema within a given domain, a typically hierarchical data structure containing all the relevant entities and their relationships and rules (theorems, regulations) within that domain.

T. R. Gruber has described an ontology in this sense as "an explicit specification of conceptualization".

Although the term 'ontology' has been used very loosely to label almost any conceptual classification scheme, among practicing computational ontologists, a true ontology should contain at a minimum not only a hierarchy of concepts organized by the subsumption relation (often called 'isa', 'subtype' or 'subclass'), but other 'semantic relations' that specify how one concept is related to another. The most common of the semantic relations other than subsumption is the 'part-of' relation. In one formal notation, one might see a relation such as (isPartOf Spine Vertebrate), meaning that a 'Spine' (in that specific sense) is part of a Vertebrate. The ontologies are organized by concepts, not words, so that the concept 'spine' referring to the spine of a book would have to be labeled by a different term, such as 'BookSpine'.

This is different from but related to the philosophical meaning of the word ontology, the study of existence. The purpose of a computational ontology is not to specify what does or does not 'exist', but to create a database, which is a human artifact, containing concepts referring to entities of interest to the ontologist, and which will be useful in performing certain types of computations. For this reason, the abstruse reasoning used by philosophical

ontologists can be helpful in recognizing and avoiding potential logical ambiguities, but where alternative ontological representations can equally well serve the pragmatic purpose of the computational ontologist, time constraints usually dictate that one choice is made and others are ignored. For certain purposes, it can be better to ignore many of the details of the objects of interest. As a result, computational ontologies developed independently for different purposes will often differ greatly from each other.

Ontologies are commonly used in artificial intelligence and knowledge representation. Computer programs can use an ontology for a variety of purposes including inductive reasoning, classification, a variety of problem solving techniques, as well as to facilitate communication and sharing of information between different systems.

An ontology which is not tied to a particular problem domain but attempts to describe general entities is known as a foundation ontology or upper ontology. Typically, more specialized schema must be created to make the data useful for real world decisions.

Such ontologies are commercially valuable, creating competition to define them. Peter Murray-Rust has claimed that this leads to "semantic and ontological warfare due to competing standards", and accordingly any standard foundation ontology is likely to be contested among commercial or political parties, each with their own idea of 'what exists' (in the philosophical sense). No one upper ontology has yet gained widespread acceptance as a de facto standard. Different organizations are attempting to define standards for specific domains. The 'Process Specification Language' (PSL) created by the National Institute for Standards and Technology (NIST) is one example.

Taxonomy

Taxonomy may refer to either a hierarchical classification of things, or the principles underlying the classification. Almost anything—animate objects, inanimate objects, places, and events—may be classified according to some taxonomic scheme.

Mathematically, a taxonomy is a tree structure of classifications for a given set of objects. At the top of this structure is a single classification—the root node—that applies to all objects. Nodes below this root are more specific classifications that apply to subsets of the total set of classified objects. So for instance in Carolus Linnaeus's Scientific classification of organisms, the root is the Organism (as this applies to all living things, it is implied rather than stated explicitly). Below this are the Kingdom, Phylum, Class, Order, Family, Genus, and Species, with various other ranks sometimes inserted.

Some have argued that the human mind naturally organizes its knowledge of the world into such systems. This view is often based on the epistemology of Immanuel Kant.

Anthropologists have observed that taxonomies are generally embedded in local cultural and social systems, and serve various social functions. Perhaps the most well-known and influential study of folk taxonomies is Emile Durkheim's The Elementary Forms of Religious Life. The theories of Kant and Durkheim also influenced Claude Levi-Strauss, the founder of anthropological structuralism. Levi-Strauss wrote two important books on taxonomies; Totemism and The Savage Mind.

Such taxonomies as those analyzed by Durkheim and Levi-Strauss are sometimes called folk taxonomies to distinguish them from scientific taxonomies that claim to be disembedded from social relations and thus objective and universal. The most well-known and widely used scientific taxonomy is Linnaean taxonomy which classifies living things and originated with Carolus Linnaeus. This taxonomic system is accessible from the article evolutionary tree.

In recent years taxonomic classification has gained support from molecular systematics, a branch of bioinformatics that employs the method of gene sequencing to construct phylogenetic trees.

Thesaurus

The word thesaurus is New Latin for treasure; coined in the early 1820's. Besides its meaning as a treasury or storehouse, it more commonly means a listing of words with similar or related meanings. For example, a book of jargon for a specialized field; or more generally a list of subject headings and cross-references used in the filing and retrieval of documents. (Or indeed papers, certificates, letters, cards, records, texts, files, articles, essays and perhaps even manuscripts.)

The first example of this genre, Roget's Thesaurus, was published in 1852, having been compiled earlier, in 1805, by Peter Roget.

Although including synonyms, entries in a thesaurus should not be taken as a list of synonyms. The entries are also designed for drawing distinctions between similar words and assisting in choosing exactly the right word. Nor does a thesaurus entry define words. That work is left to the dictionary.

In Information Technology, a thesaurus represents a database or list of semantically orthogonal topical search keys. In the field of Artificial Intelligence, a thesaurus may sometimes be referred to as an ontology.

APPENDIX C: Matching Algorithms

Fingerprint matching algorithms:

The matching algorithms calculate a similarity value of a query fingerprint and a record fingerprint. Both fingerprints must be attached to the same thesaurus.

The rank of a concept *c* in the query *q* is denoted by q_c, the rank of a concept *c* in the record fingerprint *f* is denoted by f_c. The length of a fingerprint/vector *v* is denoted by len(*v*) = sqrt(sum(v_c^2)).

All ranks are corrected before they are used in the algorithms. The correction depends on the specificity of the concept and total number of records in the collexion. The collexion *thresholds* are applied on the corrected values, on the record fingerprint as well as on the query. Additionally, the *MinimumRank* property is applied on the record fingerprint (not on the query). Thus a *MinimumRank* with a value less than *threshold* has no effect.

By default the *correction* and *threshold* parameters used on the query treatment are those from the collexion. But they can be overwritten by setting the corresponding properties on the Match object.

vector	sum(f_c * q_c) / len(f)*len(q)
portal	sum f_c, where c must be a concept given in the query q
collexis	sum(1/s_c), where c is in the fingerprint of the record and in the query q and s_c is the specificity of c
quadsum	sum(q_c^2), where c is in the fingerprint of the record
length	sqrt(sum(q_c^2))/l_q, where l_q is the length of the query fingerprint
jaccard	sum(f_c * q_c)/(len(f)² + len(q)² - sum(f_c * q_c))
dice	(2*sum(f_c * q_c)) / (len(f)² + len(q)²)
weighted	sum(f_c * q_c) * (m_f + *offset*) / (l_q + *offset*), where m_f is the number of matched concepts of f, l_q is the number concepts in q and *offset* is a given correction value.
basic	sum(f_c * q_c)

Collexis® uses the algorithms quadsum and length for internal matching - relation engine.

APPENDIX D: S&P 500 Constituents List

S&P Index 500

effective after the close April 29, 2005

S/N	Ticker	Stock Name	Sector Name	Industry Group Name
1	MMM	3M Company	Industrials	Capital Goods
2	ACE	ACE Limited	Financials	Insurance
3	ADCT	ADC Telecommunications	Information Technology	Technology Hardware & Equipment
4	AES	AES Corp.	Utilities	Utilities
5	AFL	AFLAC Inc.	Financials	Insurance
6	AT	ALLTEL Corp.	Telecommunication Services	Telecommunication Services
7	T	AT&T Corp. (New)	Telecommunication Services	Telecommunication Services
8	ABT	Abbott Labs	Health Care	Pharmaceuticals & Biotechnology
9	ADBE	Adobe Systems	Information Technology	Software & Services
10	AMD	Advanced Micro Devices	Information Technology	Semiconductors & Semiconductor Equipment
11	AET	Aetna Inc. (New)	Health Care	Health Care Equipment & Services
12	ACS	Affiliated Computer	Information Technology	Software & Services
13	A	Agilent Technologies	Information Technology	Technology Hardware & Equipment
14	APD	Air Products & Chemicals	Materials	Materials
15	ACV	Alberto-Culver	Consumer Staples	Household & Personal Products
16	ABS	Albertson's	Consumer Staples	Food & Staples Retailing
17	AA	Alcoa Inc	Materials	Materials
18	AYE	Allegheny Energy	Utilities	Utilities
19	ATI	Allegheny Technologies Inc	Materials	Materials
20	AGN	Allergan, Inc.	Health Care	Pharmaceuticals & Biotechnology
21	AW	Allied Waste Industries	Industrials	Commercial Services & Supplies
22	ALL	Allstate Corp.	Financials	Insurance
23	ALTR	Altera Corp.	Information Technology	Semiconductors & Semiconductor Equipment
24	MO	Altria Group, Inc.	Consumer Staples	Food Beverage & Tobacco

25	ASO	AmSouth Bancorporation	Financials	Banks
26	ABK	Ambac Financial Group	Financials	Insurance
27	AHC	Amerada Hess	Energy	Energy
28	AEE	Ameren Corporation	Utilities	Utilities
29	AEP	American Electric Power	Utilities	Utilities
30	AXP	American Express	Financials	Diversified Financials
31	AIG	American Int'l. Group	Financials	Insurance
32	APCC	American Power Conversion	Industrials	Capital Goods
33	ASD	American Standard	Industrials	Capital Goods
34	ABC	AmerisourceBergen Corp.	Health Care	Health Care Equipment & Services
35	AMGN	Amgen	Health Care	Pharmaceuticals & Biotechnology
36	APC	Anadarko Petroleum	Energy	Energy
37	ADI	Analog Devices	Information Technology	Semiconductors & Semiconductor Equipment
38	ANDW	Andrew Corp.	Information Technology	Technology Hardware & Equipment
39	BUD	Anheuser-Busch	Consumer Staples	Food Beverage & Tobacco
40	AOC	Aon Corp.	Financials	Insurance
41	APA	Apache Corp.	Energy	Energy
42	AIV	Apartment Investment & Mgmt'A'	Financials	Real Estate
43	APOL	Apollo Group	Consumer Discretionary	Consumer Services
44	AAPL	Apple Computer	Information Technology	Technology Hardware & Equipment
45	ABI	Applera Corp-Applied Biosystems Group	Health Care	Pharmaceuticals & Biotechnology
46	AMAT	Applied Materials	Information Technology	Semiconductors & Semiconductor Equipment
47	AMCC	Applied Micro Circuits	Information Technology	Semiconductors & Semiconductor Equipment
48	ADM	Archer-Daniels-Midland	Consumer Staples	Food Beverage & Tobacco
49	ASN	Archstone-Smith Trust	Financials	Real Estate
50	ASH	Ashland Inc.	Energy	Energy
51	AN	AutoNation, Inc.	Consumer Discretionary	Retailing
52	AZO	AutoZone Inc.	Consumer Discretionary	Retailing
53	ADSK	Autodesk, Inc.	Information Technology	Software & Services
54	ADP	Automatic Data Processing Inc.	Information Technology	Software & Services
55	AV	Avaya Inc.	Information Technology	Technology Hardware & Equipment
56	AVY	Avery Dennison Corp.	Industrials	Commercial Services & Supplies
57	AVP	Avon Products	Consumer Staples	Household & Personal Products
58	BBT	BB&T Corporation	Financials	Banks
59	BIIB	BIOGEN IDEC Inc.	Health Care	Pharmaceuticals & Biotechnology
60	BJS	BJ Services	Energy	Energy
61	BMC	BMC Software	Information Technology	Software & Services
62	BHI	Baker Hughes	Energy	Energy
63	BLL	Ball Corp.	Materials	Materials
64	BAC	Bank of America Corp.	Financials	Banks
65	BK	Bank of New York	Financials	Diversified Financials
66	BCR	Bard (C.R.) Inc.	Health Care	Health Care Equipment & Services
67	BOL	Bausch & Lomb	Health Care	Health Care Equipment & Services
68	BAX	Baxter International Inc.	Health Care	Health Care Equipment & Services
69	BSC	Bear Stearns Cos.	Financials	Diversified Financials
70	BDX	Becton, Dickinson	Health Care	Health Care Equipment & Services
71	BBBY	Bed Bath & Beyond	Consumer Discretionary	Retailing
72	BLS	BellSouth	Telecommunication Services	Telecommunication Services
73	BMS	Bemis Company	Materials	Materials
74	BBY	Best Buy Co., Inc.	Consumer Discretionary	Retailing
75	BLI	Big Lots, Inc.	Consumer Discretionary	Retailing
76	BMET	Biomet, Inc.	Health Care	Health Care Equipment & Services

77	BDK	Black & Decker Corp.	Consumer Discretionary	Consumer Durables & Apparel
78	HRB	Block H&R	Consumer Discretionary	Consumer Services
79	BA	Boeing Company	Industrials	Capital Goods
80	BSX	Boston Scientific	Health Care	Health Care Equipment & Services
81	BMY	Bristol-Myers Squibb	Health Care	Pharmaceuticals & Biotechnology
82	BRCM	Broadcom Corporation	Information Technology	Semiconductors & Semiconductor Equipment
83	BF.B	Brown-Forman Corp.	Consumer Staples	Food Beverage & Tobacco
84	BC	Brunswick Corp.	Consumer Discretionary	Consumer Durables & Apparel
85	BNI	Burlington Northern Santa Fe C	Industrials	Transportation
86	BR	Burlington Resources	Energy	Energy
87	CI	CIGNA Corp.	Health Care	Health Care Equipment & Services
88	CIN	CINergy Corp.	Utilities	Utilities
89	CIT	CIT Group	Financials	Diversified Financials
90	CMS	CMS Energy	Utilities	Utilities
91	CSX	CSX Corp.	Industrials	Transportation
92	CVS	CVS Corp.	Consumer Staples	Food & Staples Retailing
93	CPN	Calpine Corp.	Utilities	Utilities
94	CPB	Campbell Soup	Consumer Staples	Food Beverage & Tobacco
95	COF	Capital One Financial	Financials	Diversified Financials
96	CAH	Cardinal Health, Inc.	Health Care	Health Care Equipment & Services
97	CMX	Caremark Rx	Health Care	Health Care Equipment & Services
98	CCL	Carnival Corp.	Consumer Discretionary	Consumer Services
99	CAT	Caterpillar Inc.	Industrials	Capital Goods
100	CD	Cendant Corporation	Industrials	Commercial Services & Supplies
101	CNP	CenterPoint Energy	Utilities	Utilities
102	CTX	Centex Corp.	Consumer Discretionary	Consumer Durables & Apparel
103	CTL	Century Telephone	Telecommunication Services	Telecommunication Services
104	SCH	Charles Schwab	Financials	Diversified Financials
105	CVX	ChevronTexaco Corp.	Energy	Energy
106	CHIR	Chiron Corp.	Health Care	Pharmaceuticals & Biotechnology
107	CB	Chubb Corp.	Financials	Insurance
108	CIEN	Ciena Corp.	Information Technology	Technology Hardware & Equipment
109	CINF	Cincinnati Financial	Financials	Insurance
110	CTAS	Cintas Corporation	Industrials	Commercial Services & Supplies
111	CC	Circuit City Group	Consumer Discretionary	Retailing
112	CSCO	Cisco Systems	Information Technology	Technology Hardware & Equipment
113	C	Citigroup Inc.	Financials	Diversified Financials
114	CZN	Citizens Communications	Telecommunication Services	Telecommunication Services
115	CTXS	Citrix Systems	Information Technology	Software & Services
116	CCU	Clear Channel Communications	Consumer Discretionary	Media
117	CLX	Clorox Co.	Consumer Staples	Household & Personal Products
118	COH	Coach, Inc.	Consumer Discretionary	Consumer Durables & Apparel
119	KO	Coca Cola Co.	Consumer Staples	Food Beverage & Tobacco
120	CCE	Coca-Cola Enterprises	Consumer Staples	Food Beverage & Tobacco
121	CL	Colgate-Palmolive	Consumer Staples	Household & Personal Products
122	CMCSA	Comcast Corp.	Consumer Discretionary	Media
123	CMA	Comerica Inc.	Financials	Banks
124	CBSS	Compass Bancshares	Financials	Banks
125	CA	Computer Associates Intl.	Information Technology	Software & Services
126	CSC	Computer Sciences Corp.	Information Technology	Software & Services
127	CPWR	Compuware Corp.	Information Technology	Software & Services
128	CMVT	Comverse Technology	Information Technology	Technology Hardware & Equipment

129	CAG	ConAgra Foods, Inc.	Consumer Staples	Food Beverage & Tobacco
130	COP	ConocoPhillips	Energy	Energy
131	ED	Consolidated Edison	Utilities	Utilities
132	CEG	Constellation Energy Group	Utilities	Utilities
133	CVG	Convergys Corp.	Information Technology	Software & Services
134	CBE	Cooper Industries, Ltd.	Industrials	Capital Goods
135	CTB	Cooper Tire & Rubber	Consumer Discretionary	Automobiles & Components
136	GLW	Corning Inc.	Information Technology	Technology Hardware & Equipment
137	COST	Costco Co.	Consumer Staples	Food & Staples Retailing
138	CFC	Countrywide Financial Corp.	Financials	Banks
139	CMI	Cummins Inc.	Industrials	Capital Goods
140	DTE	DTE Energy Co.	Utilities	Utilities
141	DCN	Dana Corp.	Consumer Discretionary	Automobiles & Components
142	DHR	Danaher Corp.	Industrials	Capital Goods
143	DRI	Darden Restaurants	Consumer Discretionary	Consumer Services
144	DE	Deere & Co.	Industrials	Capital Goods
145	DELL	Dell Inc.	Information Technology	Technology Hardware & Equipment
146	DPH	Delphi Corporation	Consumer Discretionary	Automobiles & Components
147	DAL	Delta Air Lines	Industrials	Transportation
148	DVN	Devon Energy Corp.	Energy	Energy
149	DDS	Dillard Inc.	Consumer Discretionary	Retailing
150	DG	Dollar General	Consumer Discretionary	Retailing
151	D	Dominion Resources	Utilities	Utilities
152	RRD	Donnelley (R.R.) & Sons	Industrials	Commercial Services & Supplies
153	DOV	Dover Corp.	Industrials	Capital Goods
154	DOW	Dow Chemical	Materials	Materials
155	DJ	Dow Jones & Co.	Consumer Discretionary	Media
156	DD	Du Pont (E.I.)	Materials	Materials
157	DUK	Duke Energy	Utilities	Utilities
158	DYN	Dynegy Inc. (New) Class A	Utilities	Utilities
159	ET	E*Trade Financial Corp.	Financials	Diversified Financials
160	EMC	EMC Corp.	Information Technology	Technology Hardware & Equipment
161	EOG	EOG Resources	Energy	Energy
162	EMN	Eastman Chemical	Materials	Materials
163	EK	Eastman Kodak	Consumer Discretionary	Consumer Durables & Apparel
164	ETN	Eaton Corp.	Industrials	Capital Goods
165	ECL	Ecolab Inc.	Materials	Materials
166	EIX	Edison Int'l	Utilities	Utilities
167	EP	El Paso Corp.	Energy	Energy
168	ERTS	Electronic Arts	Information Technology	Software & Services
169	EDS	Electronic Data Systems	Information Technology	Software & Services
170	EMR	Emerson Electric	Industrials	Capital Goods
171	EC	Engelhard Corp.	Materials	Materials
172	ETR	Entergy Corp.	Utilities	Utilities
173	EFX	Equifax Inc.	Industrials	Commercial Services & Supplies
174	EOP	Equity Office Properties	Financials	Real Estate
175	EQR	Equity Residential	Financials	Real Estate
176	EXC	Exelon Corp.	Utilities	Utilities
177	ESRX	Express Scripts	Health Care	Health Care Equipment & Services
178	XOM	Exxon Mobil Corp.	Energy	Energy
179	FISV	FIserv Inc.	Information Technology	Software & Services
180	FPL	FPL Group	Utilities	Utilities

181	FDO	Family Dollar Stores	Consumer Discretionary	Retailing
182	FNM	Fannie Mae	Financials	Banks
183	FDX	FedEx Corporation	Industrials	Transportation
184	FRE	Federal Home Loan Mtg.	Financials	Banks
185	FD	Federated Dept. Stores	Consumer Discretionary	Retailing
186	FII	Federated Investors Inc.	Financials	Diversified Financials
187	FITB	Fifth Third Bancorp	Financials	Banks
188	FDC	First Data	Information Technology	Software & Services
189	FHN	First Horizon National	Financials	Banks
190	FE	FirstEnergy Corp.	Utilities	Utilities
191	FSH	Fisher Scientific	Health Care	Health Care Equipment & Services
192	FLR	Fluor Corp. (New)	Industrials	Capital Goods
193	F	Ford Motor	Consumer Discretionary	Automobiles & Components
194	FRX	Forest Laboratories	Health Care	Pharmaceuticals & Biotechnology
195	FO	Fortune Brands, Inc.	Consumer Discretionary	Consumer Durables & Apparel
196	BEN	Franklin Resources	Financials	Diversified Financials
197	FCX	Freeport-McMoran Cp & Gld	Materials	Materials
198	FSL.B	Freescale Semiconductor Inc.	Information Technology	Semiconductors & Semiconductor Equipment
199	GCI	Gannett Co.	Consumer Discretionary	Media
200	GPS	Gap (The)	Consumer Discretionary	Retailing
201	GTW	Gateway Inc.	Information Technology	Technology Hardware & Equipment
202	GD	General Dynamics	Industrials	Capital Goods
203	GE	General Electric	Industrials	Capital Goods
204	GIS	General Mills	Consumer Staples	Food Beverage & Tobacco
205	GM	General Motors	Consumer Discretionary	Automobiles & Components
206	GPC	Genuine Parts	Consumer Discretionary	Retailing
207	GENZ	Genzyme Corp.	Health Care	Pharmaceuticals & Biotechnology
208	GP	Georgia-Pacific Group	Materials	Materials
209	GILD	Gilead Sciences	Health Care	Pharmaceuticals & Biotechnology
210	G	Gillette Co.	Consumer Staples	Household & Personal Products
211	GDW	Golden West Financial	Financials	Banks
212	GS	Goldman Sachs Group	Financials	Diversified Financials
213	GR	Goodrich Corporation	Industrials	Capital Goods
214	GT	Goodyear Tire & Rubber	Consumer Discretionary	Automobiles & Components
215	GWW	Grainger (W.W.) Inc.	Industrials	Capital Goods
216	GLK	Great Lakes Chemical	Materials	Materials
217	GDT	Guidant Corp.	Health Care	Health Care Equipment & Services
218	HCA	HCA Inc.	Health Care	Health Care Equipment & Services
219	HAL	Halliburton Co.	Energy	Energy
220	HDI	Harley-Davidson	Consumer Discretionary	Automobiles & Components
221	HET	Harrah's Entertainment	Consumer Discretionary	Consumer Services
222	HIG	Hartford Financial Svc.Gp.	Financials	Insurance
223	HAS	Hasbro Inc.	Consumer Discretionary	Consumer Durables & Apparel
224	HMA	Health Management Assoc.	Health Care	Health Care Equipment & Services
225	HNZ	Heinz (H.J.)	Consumer Staples	Food Beverage & Tobacco
226	HPC	Hercules, Inc.	Materials	Materials
227	HPQ	Hewlett-Packard	Information Technology	Technology Hardware & Equipment
228	HLT	Hilton Hotels	Consumer Discretionary	Consumer Services
229	HD	Home Depot	Consumer Discretionary	Retailing
230	HON	Honeywell Int'l Inc.	Industrials	Capital Goods
231	HSP	Hospira Inc.	Health Care	Health Care Equipment & Services
232	HUM	Humana Inc.	Health Care	Health Care Equipment & Services

233	HBAN	Huntington Bancshares	Financials	Banks
234	RX	IMS Health Inc.	Health Care	Health Care Equipment & Services
235	ITT	ITT Industries, Inc.	Industrials	Capital Goods
236	ITW	Illinois Tool Works	Industrials	Capital Goods
237	IR	Ingersoll-Rand Co. Ltd.	Industrials	Capital Goods
238	INTC	Intel Corp.	Information Technology	Semiconductors & Semiconductor Equipment
239	IBM	International Bus. Machines	Information Technology	Technology Hardware & Equipment
240	IFF	International Flav/Frag	Materials	Materials
241	IGT	International Game Technology	Consumer Discretionary	Consumer Services
242	IP	International Paper	Materials	Materials
243	IPG	Interpublic Group	Consumer Discretionary	Media
244	INTU	Intuit, Inc.	Information Technology	Software & Services
245	JDSU	JDS Uniphase Corp	Information Technology	Technology Hardware & Equipment
246	JPM	JPMorgan Chase & Co.	Financials	Diversified Financials
247	JBL	Jabil Circuit	Information Technology	Technology Hardware & Equipment
248	JNS	Janus Capital Group	Financials	Diversified Financials
249	JP	Jefferson-Pilot	Financials	Insurance
250	JNJ	Johnson & Johnson	Health Care	Pharmaceuticals & Biotechnology
251	JCI	Johnson Controls	Consumer Discretionary	Automobiles & Components
252	JNY	Jones Apparel Group	Consumer Discretionary	Consumer Durables & Apparel
253	KBH	KB Home	Consumer Discretionary	Consumer Durables & Apparel
254	KLAC	KLA-Tencor Corp.	Information Technology	Semiconductors & Semiconductor Equipment
255	K	Kellogg Co.	Consumer Staples	Food Beverage & Tobacco
256	KMG	Kerr-McGee	Energy	Energy
257	KEY	KeyCorp	Financials	Banks
258	KSE	Keyspan Energy	Utilities	Utilities
259	KMB	Kimberly-Clark	Consumer Staples	Household & Personal Products
260	KMI	Kinder Morgan	Energy	Energy
261	KG	King Pharmaceuticals	Health Care	Pharmaceuticals & Biotechnology
262	KRI	Knight-Ridder Inc.	Consumer Discretionary	Media
263	KSS	Kohl's Corp.	Consumer Discretionary	Retailing
264	KR	Kroger Co.	Consumer Staples	Food & Staples Retailing
265	LLL	L-3 Communications Holdings	Industrials	Capital Goods
266	LSI	LSI Logic	Information Technology	Semiconductors & Semiconductor Equipment
267	LH	Laboratory Corp. of America Holding	Health Care	Health Care Equipment & Services
268	LEG	Leggett & Platt	Consumer Discretionary	Consumer Durables & Apparel
269	LEH	Lehman Bros.	Financials	Diversified Financials
270	LXK	Lexmark Int'l Inc	Information Technology	Technology Hardware & Equipment
271	LLY	Lilly (Eli) & Co.	Health Care	Pharmaceuticals & Biotechnology
272	LTD	Limited Brands, Inc.	Consumer Discretionary	Retailing
273	LNC	Lincoln National	Financials	Insurance
274	LLTC	Linear Technology Corp.	Information Technology	Semiconductors & Semiconductor Equipment
275	LIZ	Liz Claiborne, Inc.	Consumer Discretionary	Consumer Durables & Apparel
276	LMT	Lockheed Martin Corp.	Industrials	Capital Goods
277	LTR	Loews Corp.	Financials	Insurance
278	LPX	Louisiana Pacific	Materials	Materials
279	LOW	Lowe's Cos.	Consumer Discretionary	Retailing
280	LU	Lucent Technologies	Information Technology	Technology Hardware & Equipment
281	MTB	M&T Bank Corp.	Financials	Banks
282	MBI	MBIA Inc.	Financials	Insurance
283	KRB	MBNA Corp.	Financials	Diversified Financials
284	MTG	MGIC Investment	Financials	Banks

285	HCR	Manor Care Inc.	Health Care	Health Care Equipment & Services
286	MRO	Marathon Oil Corp.	Energy	Energy
287	MAR	Marriott Int'l.	Consumer Discretionary	Consumer Services
288	MMC	Marsh & McLennan	Financials	Insurance
289	MI	Marshall & Ilsley Corp.	Financials	Banks
290	MAS	Masco Corp.	Industrials	Capital Goods
291	MAT	Mattel, Inc.	Consumer Discretionary	Consumer Durables & Apparel
292	MXIM	Maxim Integrated Prod	Information Technology	Semiconductors & Semiconductor Equipment
293	MAY	May Dept. Stores	Consumer Discretionary	Retailing
294	MYG	Maytag Corp.	Consumer Discretionary	Consumer Durables & Apparel
295	MKC	McCormick & Co.	Consumer Staples	Food Beverage & Tobacco
296	MCD	McDonald's Corp.	Consumer Discretionary	Consumer Services
297	MHP	McGraw-Hill	Consumer Discretionary	Media
298	MCK	McKesson Corp. (New)	Health Care	Health Care Equipment & Services
299	MWV	MeadWestvaco Corporation	Materials	Materials
300	MEDI	MedImmune Inc.	Health Care	Pharmaceuticals & Biotechnology
301	MHS	Medco Health Solutions Inc.	Health Care	Health Care Equipment & Services
302	MDT	Medtronic Inc.	Health Care	Health Care Equipment & Services
303	MEL	Mellon Bank Corp.	Financials	Diversified Financials
304	MRK	Merck & Co.	Health Care	Pharmaceuticals & Biotechnology
305	MERQ	Mercury Interactive	Information Technology	Software & Services
306	MDP	Meredith Corp.	Consumer Discretionary	Media
307	MER	Merrill Lynch	Financials	Diversified Financials
308	MET	MetLife Inc.	Financials	Insurance
309	MU	Micron Technology	Information Technology	Semiconductors & Semiconductor Equipment
310	MSFT	Microsoft Corp.	Information Technology	Software & Services
311	MIL	Millipore Corp.	Health Care	Health Care Equipment & Services
312	MOLX	Molex Inc.	Information Technology	Technology Hardware & Equipment
313	TAP	Molson Coors Brewing Company	Consumer Staples	Food Beverage & Tobacco
314	MON	Monsanto Co.	Materials	Materials
315	MNST	Monster Worldwide	Industrials	Commercial Services & Supplies
316	MCO	Moody's Corp	Financials	Diversified Financials
317	MWD	Morgan Stanley	Financials	Diversified Financials
318	MOT	Motorola Inc.	Information Technology	Technology Hardware & Equipment
319	MYL	Mylan Laboratories	Health Care	Pharmaceuticals & Biotechnology
320	NCR	NCR Corp.	Information Technology	Technology Hardware & Equipment
321	GAS	NICOR Inc.	Utilities	Utilities
322	NKE	NIKE Inc.	Consumer Discretionary	Consumer Durables & Apparel
323	NVDA	NVIDIA Corp.	Information Technology	Semiconductors & Semiconductor Equipment
324	NBR	Nabors Industries Ltd.	Energy	Energy
325	NCC	National City Corp.	Financials	Banks
326	NOV	National Oilwell Varco, Inc.	Energy	Energy
327	NSM	National Semiconductor	Information Technology	Semiconductors & Semiconductor Equipment
328	NAV	Navistar International Corp.	Industrials	Capital Goods
329	NTAP	Network Appliance	Information Technology	Technology Hardware & Equipment
330	NYT	New York Times Cl. A	Consumer Discretionary	Media
331	NWL	Newell Rubbermaid Co.	Consumer Discretionary	Consumer Durables & Apparel
332	NEM	Newmont Mining Corp. (Hldg. Co.)	Materials	Materials
333	NWS.A	News Corporation	Consumer Discretionary	Media
334	NXTL	Nextel Communications	Telecommunication Services	Telecommunication Services
335	NI	NiSource Inc.	Utilities	Utilities
336	NE	Noble Corporation	Energy	Energy

337	JWN	Nordstrom	Consumer Discretionary	Retailing
338	NSC	Norfolk Southern Corp.	Industrials	Transportation
339	NFB	North Fork Bancorporation	Financials	Banks
340	NTRS	Northern Trust Corp.	Financials	Diversified Financials
341	NOC	Northrop Grumman Corp.	Industrials	Capital Goods
342	NOVL	Novell Inc.	Information Technology	Software & Services
343	NVLS	Novellus Systems	Information Technology	Semiconductors & Semiconductor Equipment
344	NUE	Nucor Corp.	Materials	Materials
345	OXY	Occidental Petroleum	Energy	Energy
346	ODP	Office Depot	Consumer Discretionary	Retailing
347	OMX	OfficeMax Inc.	Consumer Discretionary	Retailing
348	OMC	Omnicom Group	Consumer Discretionary	Media
349	ORCL	Oracle Corp.	Information Technology	Software & Services
350	PCAR	PACCAR Inc.	Industrials	Capital Goods
351	PCG	PG&E Corp.	Utilities	Utilities
352	PMCS	PMC-Sierra Inc.	Information Technology	Semiconductors & Semiconductor Equipment
353	PNC	PNC Bank Corp.	Financials	Banks
354	PPG	PPG Industries	Materials	Materials
355	PPL	PPL Corp.	Utilities	Utilities
356	PTV	Pactiv Corp.	Materials	Materials
357	PLL	Pall Corp.	Industrials	Capital Goods
358	PMTC	Parametric Technology	Information Technology	Software & Services
359	PH	Parker-Hannifin	Industrials	Capital Goods
360	PAYX	Paychex Inc.	Information Technology	Software & Services
361	JCP	Penney (J.C.)	Consumer Discretionary	Retailing
362	PGL	Peoples Energy	Utilities	Utilities
363	PBG	Pepsi Bottling Group	Consumer Staples	Food Beverage & Tobacco
364	PEP	PepsiCo Inc.	Consumer Staples	Food Beverage & Tobacco
365	PKI	PerkinElmer	Health Care	Health Care Equipment & Services
366	PFE	Pfizer, Inc.	Health Care	Pharmaceuticals & Biotechnology
367	PD	Phelps Dodge	Materials	Materials
368	PNW	Pinnacle West Capital	Utilities	Utilities
369	PBI	Pitney-Bowes	Industrials	Commercial Services & Supplies
370	PCL	Plum Creek Timber Co.	Financials	Real Estate
371	PX	Praxair, Inc.	Materials	Materials
372	PFG	Principal Financial Group	Financials	Diversified Financials
373	PLD	ProLogis	Financials	Real Estate
374	PG	Procter & Gamble	Consumer Staples	Household & Personal Products
375	PGN	Progress Energy, Inc.	Utilities	Utilities
376	PGR	Progressive Corp.	Financials	Insurance
377	PVN	Providian Financial Corp.	Financials	Diversified Financials
378	PRU	Prudential Financial	Financials	Insurance
379	PEG	Public Serv. Enterprise Inc.	Utilities	Utilities
380	PHM	Pulte Homes, Inc.	Consumer Discretionary	Consumer Durables & Apparel
381	QLGC	QLogic Corp.	Information Technology	Technology Hardware & Equipment
382	QCOM	QUALCOMM Inc.	Information Technology	Technology Hardware & Equipment
383	DGX	Quest Diagnostics	Health Care	Health Care Equipment & Services
384	Q	Qwest Communications Int	Telecommunication Services	Telecommunication Services
385	RSH	RadioShack Corp	Consumer Discretionary	Retailing
386	RTN	Raytheon Co. (New)	Industrials	Capital Goods
387	RBK	Reebok International	Consumer Discretionary	Consumer Durables & Apparel
388	RF	Regions Financial Corp. (New)	Financials	Banks

389	RAI	Reynolds American Inc.	Consumer Staples	Food Beverage & Tobacco
390	RHI	Robert Half International	Industrials	Commercial Services & Supplies
391	ROK	Rockwell Automation, Inc.	Industrials	Capital Goods
392	COL	Rockwell Collins	Industrials	Capital Goods
393	ROH	Rohm & Haas	Materials	Materials
394	RDC	Rowan Cos.	Energy	Energy
395	R	Ryder System	Industrials	Transportation
396	SAFC	SAFECO Corp.	Financials	Insurance
397	SBC	SBC Communications Inc.	Telecommunication Services	Telecommunication Services
398	SLM	SLM Corporation	Financials	Diversified Financials
399	TSG	Sabre Holding Corp.	Information Technology	Software & Services
400	SWY	Safeway Inc.	Consumer Staples	Food & Staples Retailing
401	SANM	Sanmina-SCI Corp.	Information Technology	Technology Hardware & Equipment
402	SLE	Sara Lee Corp.	Consumer Staples	Food Beverage & Tobacco
403	SGP	Schering-Plough	Health Care	Pharmaceuticals & Biotechnology
404	SLB	Schlumberger Ltd.	Energy	Energy
405	SFA	Scientific-Atlanta	Information Technology	Technology Hardware & Equipment
406	SEE	Sealed Air Corp.(New)	Materials	Materials
407	SHLD	Sears Holdings Corporation	Consumer Discretionary	Retailing
408	SRE	Sempra Energy	Utilities	Utilities
409	SHW	Sherwin-Williams	Consumer Discretionary	Retailing
410	SEBL	Siebel Systems Inc	Information Technology	Software & Services
411	SIAL	Sigma-Aldrich	Materials	Materials
412	SPG	Simon Property Group, Inc	Financials	Real Estate
413	SNA	Snap-On Inc.	Consumer Discretionary	Consumer Durables & Apparel
414	SLR	Solectron	Information Technology	Technology Hardware & Equipment
415	SO	Southern Co.	Utilities	Utilities
416	LUV	Southwest Airlines	Industrials	Transportation
417	SOV	Sovereign Bancorp	Financials	Banks
418	FON	Sprint Corp. FON	Telecommunication Services	Telecommunication Services
419	STJ	St Jude Medical	Health Care	Health Care Equipment & Services
420	STA	St. Paul Travelers Cos.	Financials	Insurance
421	SWK	Stanley Works	Consumer Discretionary	Consumer Durables & Apparel
422	SPLS	Staples Inc.	Consumer Discretionary	Retailing
423	SBUX	Starbucks Corp.	Consumer Discretionary	Consumer Services
424	HOT	Starwood Hotels & Resorts	Consumer Discretionary	Consumer Services
425	STT	State Street Corp.	Financials	Diversified Financials
426	SYK	Stryker Corp.	Health Care	Health Care Equipment & Services
427	SUNW	Sun Microsystems	Information Technology	Technology Hardware & Equipment
428	SDS	SunGard Data Systems	Information Technology	Software & Services
429	STI	SunTrust Banks	Financials	Banks
430	SUN	Sunoco, Inc.	Energy	Energy
431	SVU	Supervalu Inc.	Consumer Staples	Food & Staples Retailing
432	SYMC	Symantec Corp.	Information Technology	Software & Services
433	SBL	Symbol Technologies	Information Technology	Technology Hardware & Equipment
434	SNV	Synovus Financial	Financials	Banks
435	SYY	Sysco Corp.	Consumer Staples	Food & Staples Retailing
436	TROW	T. Rowe Price Group	Financials	Diversified Financials
437	TE	TECO Energy	Utilities	Utilities
438	TJX	TJX Companies Inc.	Consumer Discretionary	Retailing
439	TXU	TXU Corp.	Utilities	Utilities
440	TGT	Target Corp.	Consumer Discretionary	Retailing

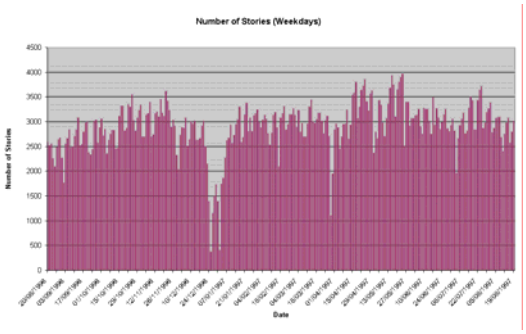
441	TEK	Tektronix Inc.	Information Technology	Technology Hardware & Equipment
442	TLAB	Tellabs, Inc.	Information Technology	Technology Hardware & Equipment
443	TIN	Temple-Inland	Materials	Materials
444	THC	Tenet Healthcare Corp.	Health Care	Health Care Equipment & Services
445	TER	Teradyne Inc.	Information Technology	Semiconductors & Semiconductor Equipment
446	TXN	Texas Instruments	Information Technology	Semiconductors & Semiconductor Equipment
447	TXT	Textron Inc.	Industrials	Capital Goods
448	HSY	The Hershey Company	Consumer Staples	Food Beverage & Tobacco
449	TMO	Thermo Electron	Health Care	Health Care Equipment & Services
450	TIF	Tiffany & Co.	Consumer Discretionary	Retailing
451	TWX	Time Warner Inc.	Consumer Discretionary	Media
452	TMK	Torchmark Corp.	Financials	Insurance
453	TOY	Toys R Us, Inc.	Consumer Discretionary	Retailing
454	RIG	Transocean Inc.	Energy	Energy
455	TRB	Tribune Co.	Consumer Discretionary	Media
456	TYC	Tyco International	Industrials	Capital Goods
457	USB	U.S. Bancorp	Financials	Banks
458	UST	UST Inc.	Consumer Staples	Food Beverage & Tobacco
459	UNP	Union Pacific	Industrials	Transportation
460	UIS	Unisys Corp.	Information Technology	Software & Services
461	UNH	United Health Group Inc.	Health Care	Health Care Equipment & Services
462	UPS	United Parcel Service	Industrials	Transportation
463	X	United States Steel Corp.	Materials	Materials
464	UTX	United Technologies	Industrials	Capital Goods
465	UVN	Univision Communications	Consumer Discretionary	Media
466	UCL	Unocal Corp.	Energy	Energy
467	UNM	UnumProvident Corp.	Financials	Insurance
468	VFC	V.F. Corp.	Consumer Discretionary	Consumer Durables & Apparel
469	VLO	Valero Energy	Energy	Energy
470	VRTS	Veritas Software	Information Technology	Software & Services
471	VZ	Verizon Communications	Telecommunication Services	Telecommunication Services
472	VIA.B	Viacom Inc.	Consumer Discretionary	Media
473	VC	Visteon Corp.	Consumer Discretionary	Automobiles & Components
474	VMC	Vulcan Materials	Materials	Materials
475	WB	Wachovia Corp. (New)	Financials	Banks
476	WMT	Wal-Mart Stores	Consumer Staples	Food & Staples Retailing
477	WAG	Walgreen Co.	Consumer Staples	Food & Staples Retailing
478	DIS	Walt Disney Co.	Consumer Discretionary	Media
479	WM	Washington Mutual	Financials	Banks
480	WMI	Waste Management Inc.	Industrials	Commercial Services & Supplies
481	WAT	Waters Corporation	Health Care	Health Care Equipment & Services
482	WPI	Watson Pharmaceuticals	Health Care	Pharmaceuticals & Biotechnology
483	WLP	WellPoint Inc.	Health Care	Health Care Equipment & Services
484	WFC	Wells Fargo	Financials	Banks
485	WEN	Wendy's International	Consumer Discretionary	Consumer Services
486	WY	Weyerhaeuser Corp.	Materials	Materials
487	WHR	Whirlpool Corp.	Consumer Discretionary	Consumer Durables & Apparel
488	WMB	Williams Cos.	Energy	Energy
489	WWY	Wrigley (Wm) Jr.	Consumer Staples	Food Beverage & Tobacco
490	WYE	Wyeth	Health Care	Pharmaceuticals & Biotechnology
491	XL	XL Capital	Financials	Insurance
492	XTO	XTO Energy Inc.	Energy	Energy

493	XEL	Xcel Energy Inc	Utilities	Utilities
494	XRX	Xerox Corp.	Information Technology	Technology Hardware & Equipment
495	XLNX	Xilinx, Inc	Information Technology	Semiconductors & Semiconductor Equipment
496	YHOO	Yahoo Inc.	Information Technology	Software & Services
497	YUM	Yum! Brands, Inc	Consumer Discretionary	Consumer Services
498	ZMH	Zimmer Holdings	Health Care	Health Care Equipment & Services
499	ZION	Zions Bancorp	Financials	Banks
500	EBAY	eBay Inc.	Consumer Discretionary	Retailing

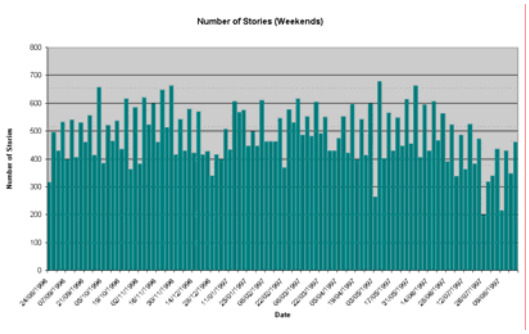
APPENDIX E: Reuters Corpus, Volume I

Statistics

Number of Stories on Weekdays

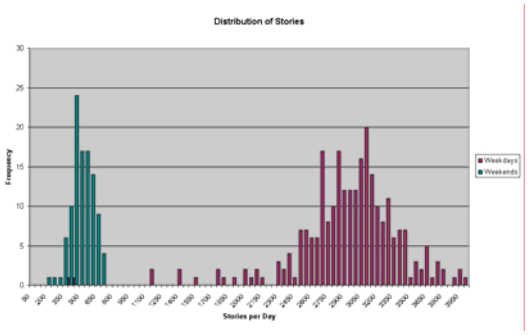


Number of Stories over the Weekend



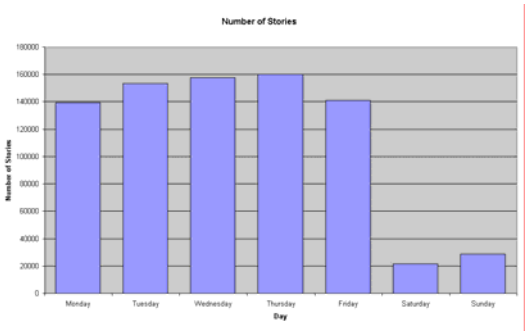
These charts shows the number of stories per day, over the period of the Corpus, since the weekend has a dramatically different number of stories this is shown on a separate chart.

Distribution of Stories



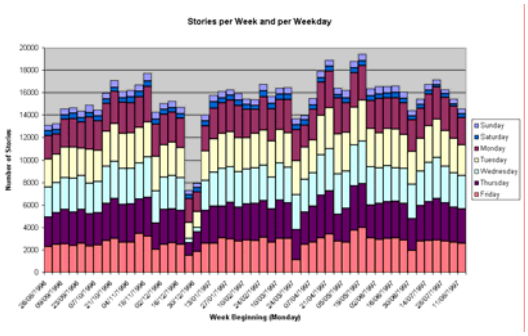
The difference between the weekdays and weekend can be shown more clearly in the graph above, showing the number of days where a particular number of stories were produced. The distribution of stories is clearly bi-modal, with a much higher median for weekdays than for weekends. This is a typical pattern in the News Industry.

Number of Stories per Day of Week



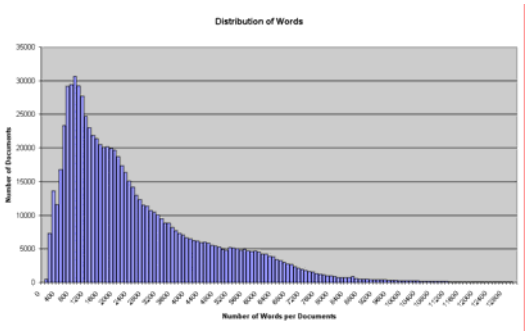
In addition to the difference between the weekdays and weekend there is a more subtle difference between the weekdays themselves. This pattern can be seen in the chart above.

Stories per Week and Weekday

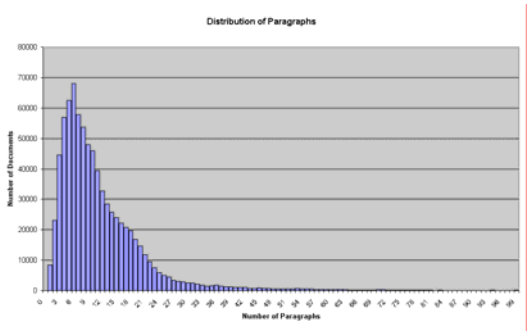


Information about the number of stories per day is summarised in the chart above. In addition to the previous comments it is clear that the weeks beginning 21st December 1996 and 30th December 1996 had a particularly low number of stories produced, especially on the New Years day and Christmas day (both falling on a Wednesday.

Number of Words per Document

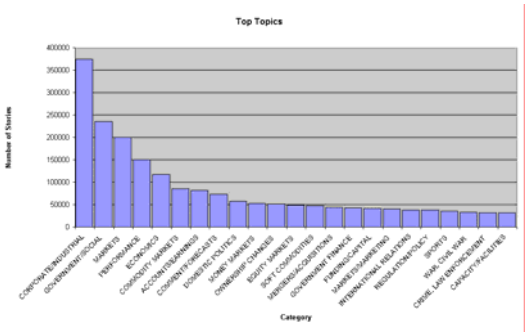


Number of Paragraphs per Document

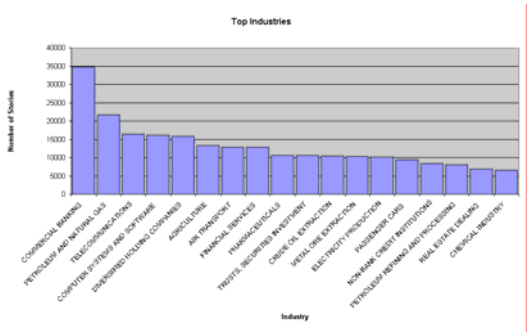


These charts show the number of stories that have a particular word or paragraph count. It can be seen that most stories are quite short, with around 6-7 paragraphs and 1000 words

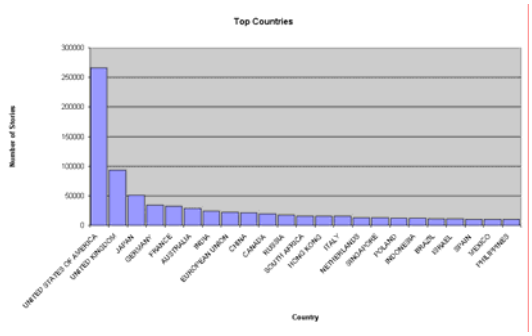
Top Topic Codes



Top Industry Codes

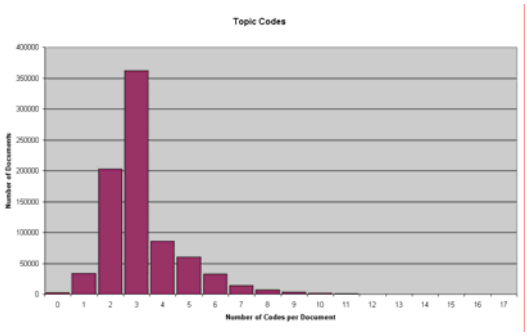


Top Country Codes

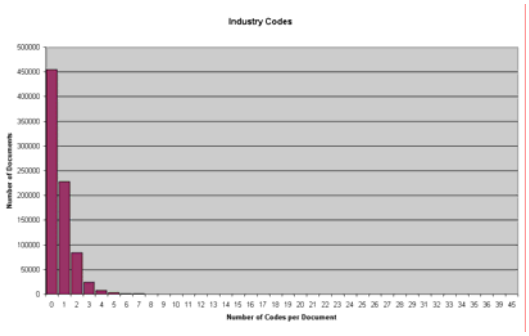


Each story is tagged with zero or more topics, zero or more countries and zero or more industries. The charts above show the frequency of occurrence of these codes. If there are more than one codes of a given category in a document then each will be counted separately.

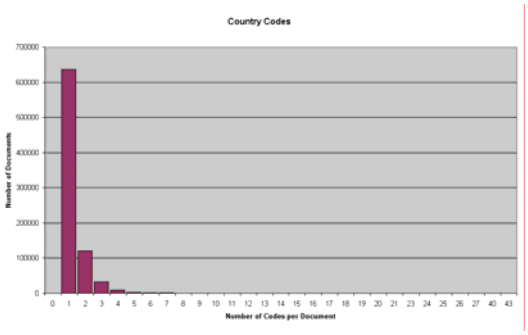
Topic Code Count



Industry Code Count

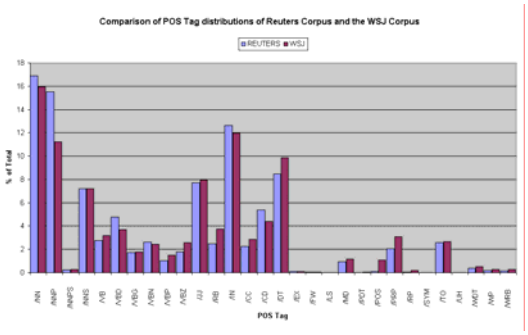


Country Code Count



As mentioned above each story can have zero or more topic, industry or country codes. The charts above show the number of documents with a particular number of codes, for each of the code types.

POS Tag Distribution



This chart shows the Part of Speech tag distributions over the Penn-Treebank tag set for both the Wall Street Journal and the Reuters news corpus. The POS tagging was carried out using Brill's tagger, freely available from Eric Brill's Homepage. As one would expect, the distributions are similar.

APPENDIX F: NewsML

An example of a Newsitem in XML

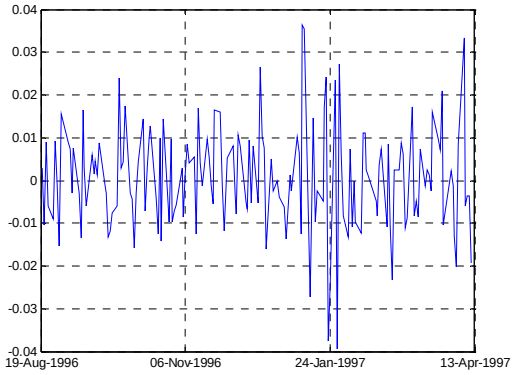
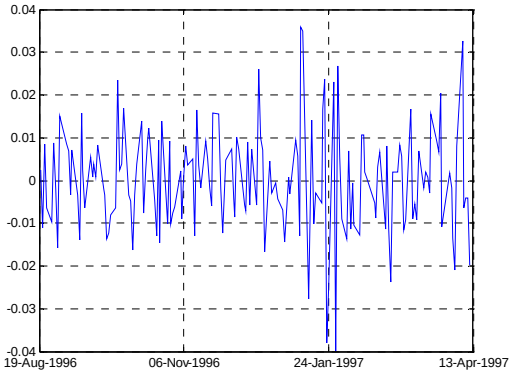
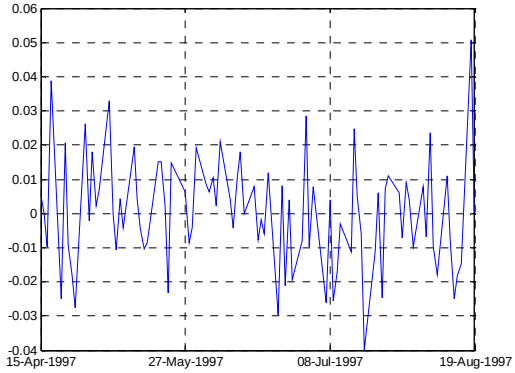
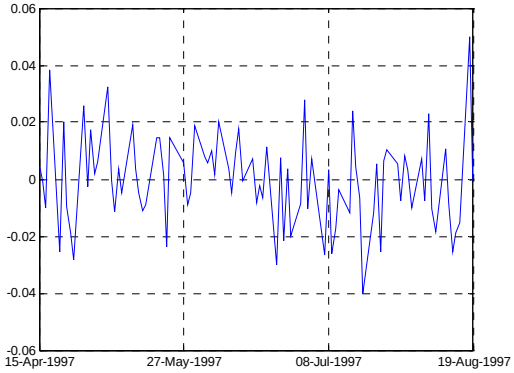
<?xml version="1.0" encoding="iso-8859-1" ?>
<newsitem itemid="40730" id="root" date="1996-09-09" xml:lang="en">
<title>USA: Sudbury rejects Park-Ohio purchase offer.</title>
<headline>Sudbury rejects Park-Ohio purchase offer.</headline>
<dateline>CLEVELAND 1996-09-09</dateline>
<text>
Sudbury Inc said Monday its board of directors had rejected as inadequate an unsolicited offer from Park-Ohio Industries Inc to buy the company for $170 million.
Sudbury said it would continue to explore ways to maximize shareholder value. Since July, Alex Brown & Sons Inc has been looking for contracts with entities interested in buying the company.
Nasdaq trading in Sudbury stock was halted at $12 3/4 Monday.
Park-Ohio shares were up 1/4 at 14-7/8 early Monday afternoon.
Park-Ohio last week proposed to buy the company for $170 million, or $11 a share.
Sudbury provides iron, aluminum and zinc castings and other metal products for appliance and automobile manufacturers, among others.
</text>
<copyright>(c) Reuters Limited 1996</copyright>

<metadata>
<codes class="bip:countries:1.0">
<code code="USA">
<editdetail attribution="Reuters BIP Coding Group" action="confirmed" date="1996-09-09"/>
</code>
</codes>
<codes class="bip:industries:1.0">
<code code="I13000">
<editdetail attribution="Reuters BIP Coding Group" action="confirmed" date="1996-09-09"/>
</code>
<code code="I1300003">
<editdetail attribution="Reuters BIP Coding Group" action="confirmed" date="1996-09-09"/>
</code>
<code code="I83960">
<editdetail attribution="Reuters BIP Coding Group" action="confirmed" date="1996-09-09"/>
</code>
</codes>
<codes class="bip:topics:1.0">
<code code="C18">
<editdetail attribution="Reuters BIP Coding Group" action="confirmed" date="1996-09-09"/>
</code>
<code code="C181">
<editdetail attribution="Reuters BIP Coding Group" action="confirmed" date="1996-09-09"/>
</code>
<code code="CCAT">
<editdetail attribution="Reuters BIP Coding Group" action="confirmed" date="1996-09-09"/>
</code>
</codes>
<dc element="dc.date.created" value="1996-09-09"/>
<dc element="dc.publisher" value="Reuters Holdings Plc"/>
<dc element="dc.date.published" value="1996-09-09"/>
<dc element="dc.source" value="Reuters"/>
<dc element="dc.creator.location" value="CLEVELAND"/>
<dc element="dc.creator.location.country.name" value="USA"/>
<dc element="dc.source" value="Reuters"/>
</metadata>
</newsitem>

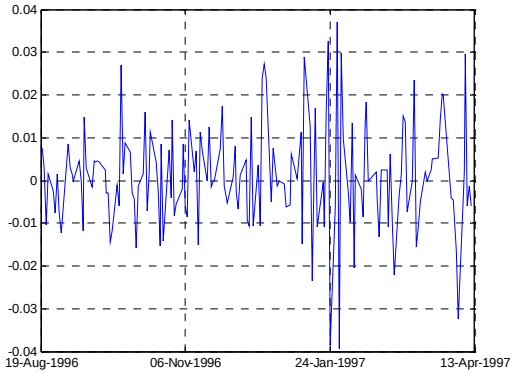
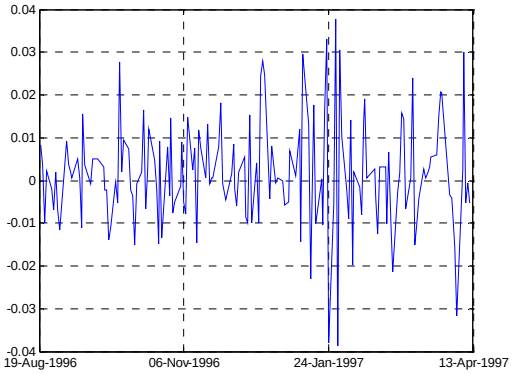
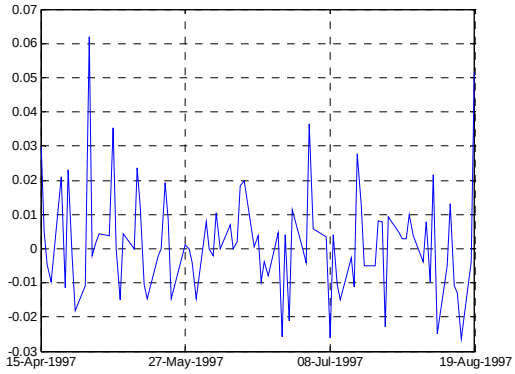
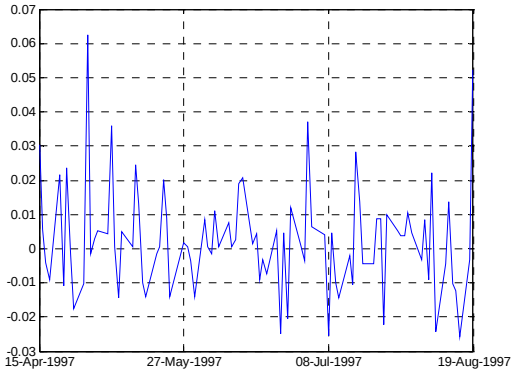
APPENDIX G: Market Dataset

Exxon Mobil Corp. Returns

Open-to-Close

Normal returns	Abnormal returns
	
	

Open-to-Open

Normal returns	Abnormal returns
	
	

APPENDIX H: Reuters Industry Codes

Industry Codes

CODE	DESCRIPTION
I0	AGRICULTURE, FORESTRY AND FISHING
I00	AGRICULTURE, FORESTRY AND FISHING
I000	AGRICULTURE, FORESTRY AND FISHING
I0000	AGRICULTURE, FORESTRY AND FISHING
I00000	AGRICULTURE, FORESTRY AND FISHING
I01	AGRICULTURE AND HORTICULTURE
I010	AGRICULTURE AND HORTICULTURE
I0100	AGRICULTURE AND HORTICULTURE
I01000	AGRICULTURE AND HORTICULTURE
I01001	AGRICULTURE
I0100105	CATTLE FARMING
I0100107	EGG PRODUCTION
I0100119	PIG FARMING
I0100121	POULTRY FARMING
I0100124	SHEEP FARMING
I0100128	SUGAR BEET GROWING
I0100132	GRAIN, CEREALS FARMING
I0100136	SUGAR CANE GROWING
I0100137	COFFEE GROWING
I0100138	COTTON GROWING
I0100141	RUBBER GROWING
I0100142	TEA GROWING
I0100144	COCOA GROWING
I0100145	TOBACCO GROWING
I01002	HORTICULTURE
I0100206	FRUIT GROWING
I0100216	VEGETABLE GROWING
I0100223	SOYA GROWING
I02	FORESTRY
I020	FORESTRY
I0200	FORESTRY
I02000	FORESTRY
I03	FISHING
I030	FISHING
I0300	FISHING
I03000	FISHING
I1	ENERGY AND WATER SUPPLY
I10	ENERGY AND WATER SUPPLY
I100	ENERGY AND WATER SUPPLY
I1000	ENERGY AND WATER SUPPLY
I10000	ENERGY AND WATER SUPPLY
I11	COAL MINING
I110	COAL MINING
I1100	COAL MINING
I11000	COAL MINING
I12	COKE PRODUCTION
I120	COKE PRODUCTION
I1200	COKE PRODUCTION
I12000	COKE PRODUCTION
I13	***PETROLEUM AND NATURAL GAS***
I130	**PETROLEUM AND NATURAL GAS**
I1300	**PETROLEUM AND NATURAL GAS**
I13000	**PETROLEUM AND NATURAL GAS**
I1300002	**CRUDE OIL EXPLORATION**
I1300003	**CRUDE OIL EXTRACTION**
I1300013	**NATURAL GAS EXPLORATION**
I1300014	**NATURAL GAS EXTRACTION**
I14	***PETROLEUM REFINING AND PROCESSING***
I140	**PETROLEUM REFINING AND PROCESSING**
I1400	**PETROLEUM REFINING AND PROCESSING**
I14000	**PETROLEUM REFINING AND PROCESSING**
I15	NUCLEAR FUEL
I150	NUCLEAR FUEL
I1500	NUCLEAR FUEL
I15000	NUCLEAR FUEL
I16	***ELECTRICITY, GAS UTILITIES***
I160	***ELECTRICITY, GAS UTILITIES***
I1600	**ELECTRICITY, GAS UTILITIES**
I16000	**ELECTRICITY, GAS UTILITIES**
I161	ELECTRICITY PRODUCTION
I1610	ELECTRICITY PRODUCTION
I16100	ELECTRICITY PRODUCTION
I16101	ELECTRICITY PRODUCTION
I1610107	HYDRO ELECTRIC POWER
I1610109	NUCLEAR POWER
I162	***GAS PRODUCTION***
I1620	**GAS PRODUCTION**
I16200	**GAS PRODUCTION**
I163	ALTERNATIVE ENERGY PRODUCTION
I1630	ALTERNATIVE ENERGY
I16300	ALTERNATIVE ENERGY
I17	WATER SUPPLY
I170	WATER SUPPLY
I1700	WATER SUPPLY

CODE	DESCRIPTION
I17000	WATER SUPPLY
I2	METALS AND MINERALS
I20	METALS AND MINERALS
I200	METALS AND MINERALS
I2000	METALS AND MINERALS
I20000	METALS AND MINERALS
I21	METAL ORE EXTRACTION
I210	METAL ORE EXTRACTION
I2100	METAL ORE EXTRACTION
I21000	METAL ORE EXTRACTION
I22	METAL MANUFACTURING
I220	METAL MANUFACTURING
I2200	METAL MANUFACTURING
I22000	METAL MANUFACTURING
I221	IRON AND STEEL
I2210	IRON AND STEEL
I22100	IRON AND STEEL
I222	STEEL TUBES
I2220	STEEL TUBES
I22200	STEEL TUBES
I223	STEEL COLD ROLLING AND FORMING
I2230	STEEL COLD ROLLING AND FORMING
I22300	STEEL COLD ROLLING AND FORMING
I224	NON FERROUS METALS
I2240	NON FERROUS METALS
I22400	NON FERROUS METALS
I2245	ALUMINIUM AND ALUMINIUM ALLOYS
I22450	ALUMINIUM AND ALUMINIUM ALLOYS
I2246	COPPER AND COPPER ALLOYS
I22460	COPPER AND COPPER ALLOYS
I2247	NON FERROUS METALS
I22470	NON FERROUS METALS
I22471	BASE NON FERROUS METALS
I22472	PRECIOUS METALS
I23	EXTRACTION OF MINERALS
I230	EXTRACTION OF MINERALS
I2300	EXTRACTION OF MINERALS
I23000	EXTRACTION OF MINERALS
I24	MINERAL PRODUCTS
I240	MINERAL PRODUCTS
I2400	MINERAL PRODUCTS
I24000	MINERAL PRODUCTS
I241	STRUCTURAL CLAY PRODUCTS
I2410	STRUCTURAL CLAY PRODUCTS
I24100	STRUCTURAL CLAY PRODUCTS
I242	CEMENT, LIME AND PLASTER
I2420	CEMENT, LIME AND PLASTER
I24200	CEMENT, LIME AND PLASTER
I243	BUILDING PRODUCTS
I2430	BUILDING PRODUCTS
I24300	BUILDING PRODUCTS
I244	ASBESTOS PRODUCTS
I2440	ASBESTOS PRODUCTS
I24400	ASBESTOS PRODUCTS
I245	STONE AND SLATE PRODUCTS
I2450	STONE AND SLATE PRODUCTS
I24500	STONE AND SLATE PRODUCTS
I246	ABRASIVE PRODUCTS
I2460	ABRASIVE PRODUCTS
I24600	ABRASIVE PRODUCTS
I247	GLASS AND GLASS PRODUCTS
I2470	GLASS AND GLASS PRODUCTS
I24700	GLASS AND GLASS PRODUCTS
I2479	GLASS AND GLASS PRODUCTS
I24794	GLASS FIBRE
I248	CERAMIC PRODUCTS
I2480	CERAMIC PRODUCTS
I24800	CERAMIC PRODUCTS
I25	CHEMICAL INDUSTRY
I250	CHEMICAL INDUSTRY
I2500	CHEMICAL INDUSTRY
I25000	CHEMICAL INDUSTRY
I251	BASIC INDUSTRIAL CHEMICALS
I2510	BASIC INDUSTRIAL CHEMICALS
I25100	BASIC INDUSTRIAL CHEMICALS

CODE	DESCRIPTION
I2511	INORGANIC CHEMICALS
I25110	INORGANIC CHEMICALS
I2512	***ORGANIC CHEMICALS INCL. PETROCHEMICALS***
I25120	**ORGANIC CHEMICALS INCL. PETROCHEMICALS**
I2513	FERTILISERS
I25130	FERTILISERS
I2514	SYNTHETIC RESINS, POLYMERS AND PLASTICS
I25140	SYNTHETIC RESINS, POLYMERS AND PLASTICS
I2516	DYESTUFFS AND PIGMENTS
I25160	DYESTUFFS AND PIGMENTS
I255	PAINTS AND INKS
I2551	PAINTS
I25510	PAINTS
I2552	PRINTING INKS
I25520	PRINTING INKS
I256	CHEMICAL INDUSTRY
I2562	ADHESIVES
I25620	ADHESIVES
I2565	EXPLOSIVES
I25650	EXPLOSIVES
I2567	***INDUSTRIAL CHEMICALS INCL. GASES***
I25670	**INDUSTRIAL CHEMICALS INCL. GASES**
I2568	PESTICIDES
I25680	PESTICIDES
I257	PHARMACEUTICALS
I2570	PHARMACEUTICALS
I25700	PHARMACEUTICALS
I258	SOAPS AND COSMETICS
I2580	SOAPS AND COSMETICS
I25800	SOAPS AND COSMETICS
I26	SYNTHETIC FIBRES INCL. POLYESTER
I260	SYNTHETIC FIBRES INCL. POLYESTER
I2600	SYNTHETIC FIBRES INCL. POLYESTER
I26000	SYNTHETIC FIBRES INCL. POLYESTER
I3	METAL GOODS AND ENGINEERING
I30	METAL GOODS AND ENGINEERING
I300	METAL GOODS AND ENGINEERING
I3000	METAL GOODS AND ENGINEERING
I30000	METAL GOODS AND ENGINEERING
I31	METAL GOODS INCL. NUTS AND BOLTS
I310	METAL GOODS INCL. NUTS AND BOLTS
I3100	METAL GOODS INCL. NUTS AND BOLTS
I31000	METAL GOODS INCL. NUTS AND BOLTS
I32	MECHANICAL ENGINEERING
I320	MECHANICAL ENGINEERING
I3200	MECHANICAL ENGINEERING
I32000	MECHANICAL ENGINEERING
I3204	INDUSTRIAL PLANT AND STEELWORK
I32040	INDUSTRIAL PLANT AND STEELWORK
I3205	BOILERS AND PROCESS PLANT FABRICATIONS
I32050	BOILERS AND PROCESS PLANT FABRICATIONS
I321	AGRICULTURAL MACHINERY
I3210	AGRICULTURAL MACHINERY
I32100	AGRICULTURAL MACHINERY
I322	MACHINE TOOLS
I3220	MACHINE TOOLS
I32200	MACHINE TOOLS
I32220	TEMPORARY
I323	TEXTILE MACHINERY
I3230	TEXTILE MACHINERY
I32300	TEXTILE MACHINERY
I324	MECHANICAL ENGINEERING
I3244	FOOD PROCESSING MACHINERY
I32440	FOOD PROCESSING MACHINERY
I3245	CHEMICAL INDUSTRY MACHINERY
I32450	CHEMICAL INDUSTRY MACHINERY
I325	MECHANICAL ENGINEERING
I3251	MINING MACHINERY
I32510	MINING MACHINERY

CODE	DESCRIPTION
I3254	CONSTRUCTION MACHINERY
I32540	CONSTRUCTION MACHINERY
I3255	LIFTING AND HANDLING EQUIPMENT
I32550	LIFTING AND HANDLING EQUIPMENT
I326	POWER TRANSMISSION EQUIPMENT
I3260	POWER TRANSMISSION EQUIPMENT
I32600	POWER TRANSMISSION EQUIPMENT
I327	MECHANICAL ENGINEERING
I32700	TEMPORARY
I3275	MECHANICAL ENGINEERING
I32751	WOODWORKING MACHINERY
I32752	RUBBER AND PLASTICS WORKING MACHINERY
I32753	LEATHERWORKING AND FOOTWEAR MACHINERY
I32754	PAPER MAKING MACHINERY
I32755	GLASS AND CERAMIC WORKING MACHINERY
I328	MECHANICAL ENGINEERING
I3281	INDUSTRIAL AND MARINE ENGINES
I32810	INDUSTRIAL AND MARINE ENGINES
I3283	COMPRESSORS, HYDRAULIC, PNEUMATIC
I32830	COMPRESSORS, HYDRAULIC, PNEUMATIC
I3284	REFRIGERATING, HEATING, AIR CONDITIONING
I32840	REFRIGERATING, HEATING, AIR CONDITIONING
I3285	MECHANICAL ENGINEERING
I32851	WEIGHING MACHINERY
I32852	PORTABLE POWER TOOLS
I3287	PUMPS
I32870	PUMPS
I329	ARMS MANUFACTURING
I3290	ARMS MANUFACTURING
I32900	ARMS MANUFACTURING
I33	METAL GOODS AND ENGINEERING
I330	METAL GOODS AND ENGINEERING
I33000	TEMPORARY
I3301	ELECTRONIC OFFICE EQUIPMENT
I33010	ELECTRONIC OFFICE EQUIPMENT
I3302	COMPUTER SYSTEMS AND SOFTWARE
I33020	COMPUTER SYSTEMS AND SOFTWARE
I3302003	COMPUTERS AND PERSONAL COMPUTERS
I3302004	COMPUTER PERIPHERALS INCL. MODEMS
I3302013	COMPUTER PRINTERS
I3302015	OPTICAL CHARACTER READERS
I3302017	COMPUTER TERMINALS, WORK STATIONS
I3302018	WORD PROCESSORS
I3302019	SECONDARY MEMORY STORES
I330202	SYSTEMS SOFTWARE
I3302020	SYSTEMS SOFTWARE
I3302021	APPLICATIONS SOFTWARE
I3302022	KNOWLEDGE BASED SYSTEMS
I3303	DATACOMMUNICATIONS AND NETWORKING
I33030	DATACOMMUNICATIONS AND NETWORKING
I34	ELECTRICAL AND ELECTRONIC ENGINEERING
I340	ELECTRICAL AND ELECTRONIC ENGINEERING
I3400	ELECTRICAL AND ELECTRONIC ENGINEERING
I34000	ELECTRICAL AND ELECTRONIC ENGINEERING
I341	WIRES AND CABLES
I3410	WIRES AND CABLES
I34100	WIRES AND CABLES
I342	BASIC ELECTRICAL EQUIPMENT
I3420	BASIC ELECTRICAL EQUIPMENT
I34200	BASIC ELECTRICAL EQUIPMENT
I343	ELECTRICAL AND ELECTRONIC ENGINEERING
I3432	BATTERIES
I34320	BATTERIES
I3433	ALARMS AND SIGNALLING EQUIPMENT
I34330	ALARMS AND SIGNALLING EQUIPMENT

CODE	DESCRIPTION
I3434	ELECTRICAL EQUIPMENT FOR VEHICLES
I34340	ELECTRICAL EQUIPMENT FOR VEHICLES
I3435	INDUSTRIAL ELECTRICAL EQUIPMENT
I34350	INDUSTRIAL ELECTRICAL EQUIPMENT
I344	TELECOMMUNICATIONS EQUIPMENT
I3440	TELECOMMUNICATIONS EQUIPMENT
I34400	TELECOMMUNICATIONS EQUIPMENT
I3441	TELEPHONE EQUIPMENT
I34410	TELEPHONE EQUIPMENT
I3442	ELECTRICAL INSTUMENTS, CONTROL SYSTEMS
I34420	ELECTRICAL INSTUMENTS, CONTROL SYSTEMS
I3443	RADAR AND ELECTRONIC EQUIPMENT
I34430	RADAR AND ELECTRONIC EQUIPMENT
I3444	PASSIVE COMPONENTS, PRINTED CIRCUITS
I34440	PASSIVE COMPONENTS, PRINTED CIRCUITS
I345	ELECTRICAL AND ELECTRONIC ENGINEERING
I3450	TEMPORARY
I3452	RECORDS, TAPES AND DISCS
I34520	RECORDS, TAPES AND DISCS
I3453	ELECTRICAL AND ELECTRONIC ENGINEERING
I34531	ELECTRONIC ACTIVE COMPONENTS
I34532	COMPONENTS FOR ELECTRONIC GOODS
I3454	ELECTRONIC CONSUMER GOODS
I34540	ELECTRONIC CONSUMER GOODS
I346	ELECTRICAL APPLIANCES
I3460	ELECTRICAL APPLIANCES
I34600	ELECTRICAL APPLIANCES
I347	ELECTRIC LIGHTING EQUIPMENT
I3470	ELECTRIC LIGHTING EQUIPMENT
I34700	ELECTRIC LIGHTING EQUIPMENT
I35	MOTOR VEHICLES AND PARTS
I350	MOTOR VEHICLES AND PARTS
I3500	MOTOR VEHICLES AND PARTS
I35000	MOTOR VEHICLES AND PARTS
I351	MOTOR VEHICLES
I3510	MOTOR VEHICLES
I35101	PASSENGER CARS
I35102	COMMERCIAL VEHICLES
I352	VEHICLE BODIES AND TRAILERS
I3520	VEHICLE BODIES AND TRAILERS
I35200	VEHICLE BODIES AND TRAILERS
I353	VEHICLE PARTS
I3530	VEHICLE PARTS
I35300	VEHICLE PARTS
I36	METAL GOODS AND ENGINEERING
I361	METAL GOODS AND ENGINEERING
I3610	METAL GOODS AND ENGINEERING
I36100	TEMPORARY
I36101	SHIPBUILDING
I36102	PLEASURE BOATS AND YACHTS
I36103	SHIPBREAKING
I362	RAILWAY EQUIPMENT
I3620	RAILWAY EQUIPMENT
I36200	RAILWAY EQUIPMENT
I363	CYCLES AND MOTOR CYCLES
I3630	CYCLES AND MOTOR CYCLES
I36300	CYCLES AND MOTOR CYCLES
I364	AEROSPACE
I3640	AEROSPACE
I36400	AEROSPACE
I3640002	AERO-ENGINES
I3640007	AIRCRAFT COMPONENTS, NOT ELECTRICAL
I364001	CIVIL AIRCRAFT MANUFACTURING
I3640010	CIVIL AIRCRAFT MANUFACTURING
I3640026	GUIDED WEAPONS
I3640029	CIVIL HELICOPTER MANUFACTURING
I364003	HOVERCRAFT MANUFACTURING
I3640030	HOVERCRAFT MANUFACTURING
I3640045	SATELLITE MANUFACTURING
I3640046	SPACECRAFT MANUFACTURING
I3640047	MILITARY AEROSPACE EQUIPMENT

CODE	DESCRIPTION
I3640048	AIRCRAFT MAINTENANCE
I37	INSTRUMENT ENGINEERING
I370	INSTRUMENT ENGINEERING
I3700	INSTRUMENT ENGINEERING
I37000	INSTRUMENT ENGINEERING
I371	MEASURING AND PRECISION INSTRUMENTS
I3710	MEASURING AND PRECISION INSTRUMENTS
I37100	MEASURING AND PRECISION INSTRUMENTS
I372	MEDICAL EQUIPMENT
I3720	MEDICAL EQUIPMENT
I37200	MEDICAL EQUIPMENT
I373	OPTICAL EQUIPMENT
I3730	OPTICAL EQUIPMENT
I37300	OPTICAL EQUIPMENT
I3733	PHOTOGRAPHIC EQUIPMENT
I37330	PHOTOGRAPHIC EQUIPMENT
I374	TIMEKEEPING EQUIPMENT
I3740	TIMEKEEPING EQUIPMENT
I37400	TIMEKEEPING EQUIPMENT
I4	PROCESSING INDUSTRIES
I40	PROCESSING INDUSTRIES
I400	PROCESSING INDUSTRIES
I4000	PROCESSING INDUSTRIES
I40000	PROCESSING INDUSTRIES
I41	FOOD, DRINK AND TOBACCO PROCESSING
I410	FOOD, DRINK AND TOBACCO PROCESSING
I4100	FOOD, DRINK AND TOBACCO PROCESSING
I41000	FOOD, DRINK AND TOBACCO PROCESSING
I411	OILS AND FATS PROCESSING
I4110	OILS AND FATS PROCESSING
I41100	OILS AND FATS PROCESSING
I412	SLAUGHTERHOUSES AND MEAT PROCESSING
I4120	SLAUGHTERHOUSES AND MEAT PROCESSING
I41200	SLAUGHTERHOUSES AND MEAT PROCESSING
I4122	MEAT PROCESSING
I41220	MEAT PROCESSING
I4123	POULTRY SLAUGHTER AND PROCESSING
I41230	POULTRY SLAUGHTER AND PROCESSING
I413	DAIRY PRODUCTS PROCESSING
I4130	DAIRY PRODUCTS PROCESSING
I41300	DAIRY PRODUCTS PROCESSING
I414	FRUIT AND VEGETABLE PROCESSING
I4140	FRUIT AND VEGETABLE PROCESSING
I41400	FRUIT AND VEGETABLE PROCESSING
I415	FISH PROCESSING
I4150	FISH PROCESSING
I41500	FISH PROCESSING
I416	GRAIN MILLING AND PROCESSING
I4160	GRAIN MILLING AND PROCESSING
I41600	GRAIN MILLING AND PROCESSING
I418	STARCH PROCESSING
I4180	STARCH PROCESSING
I41800	STARCH PROCESSING
I419	BREAD AND BISCUIT MAKING
I4190	BREAD AND BISCUIT MAKING
I41900	BREAD AND BISCUIT MAKING
I42	SUGAR AND SUGAR PRODUCTS
I420	SUGAR AND SUGAR PRODUCTS
I4200	SUGAR AND SUGAR PRODUCTS
I42000	SUGAR AND SUGAR PRODUCTS
I421	CONFECTIONERY
I4210	CONFECTIONERY
I42100	CONFECTIONERY
I422	ANIMAL FEED
I4221	ANIMAL FEED
I42210	ANIMAL FEED
I4222	PET FOODS
I42220	PET FOODS
I423	MISCELLANEOUS FOODS
I4239	MISCELLANEOUS FOODS
I42390	MISCELLANEOUS FOODS
I424	SPIRITS DISTILLING
I4240	SPIRITS DISTILLING
I42400	SPIRITS DISTILLING
I426	WINE PRODUCTION
I4260	WINE PRODUCTION
I42600	WINE PRODUCTION
I427	BREWING
I4270	BREWING
I42700	BREWING
I428	SOFT DRINKS
I4280	SOFT DRINKS
I42800	SOFT DRINKS
I429	TOBACCO
I4290	TOBACCO
I42900	TOBACCO
I43	TEXTILE INDUSTRY
I430	TEXTILE INDUSTRY
I4300	TEXTILE INDUSTRY
I43000	TEXTILE INDUSTRY
I44	LEATHER AND LEATHER GOODS
I440	LEATHER AND LEATHER GOODS
I4400	LEATHER AND LEATHER GOODS
I44000	LEATHER AND LEATHER GOODS
I45	FOOTWEAR AND CLOTHING
I450	FOOTWEAR AND CLOTHING
I4500	FOOTWEAR AND CLOTHING
I45000	FOOTWEAR AND CLOTHING
I451	FOOTWEAR
I4510	FOOTWEAR
I45100	FOOTWEAR
I453	CLOTHING
I4530	CLOTHING
I45300	CLOTHING
I455	HOUSEHOLD TEXTILES
I4550	HOUSEHOLD TEXTILES
I45500	HOUSEHOLD TEXTILES
I456	FUR GOODS
I4560	FUR GOODS
I45600	FUR GOODS
I46	TIMBER PROCESSING AND WOOD ARTICLES
I460	TIMBER PROCESSING AND WOOD ARTICLES
I4600	TIMBER PROCESSING AND WOOD ARTICLES
I46000	TIMBER PROCESSING AND WOOD ARTICLES
I467	WOODEN FURNITURE
I4670	WOODEN FURNITURE
I46700	WOODEN FURNITURE
I47	PAPER, PRINTING AND PUBLISHING
I470	PAPER, PRINTING AND PUBLISHING
I4700	PAPER, PRINTING AND PUBLISHING
I47000	PAPER, PRINTING AND PUBLISHING
I471	PAPER MANUFACTURING
I4710	PAPER MANUFACTURING
I47100	PAPER MANUFACTURING
I47101	PULP
I472	PAPER AND BOARD CONVERSION
I4720	PAPER AND BOARD CONVERSION
I47200	PAPER AND BOARD CONVERSION
I475	PRINTING AND PUBLISHING
I4750	PRINTING AND PUBLISHING
I47500	PRINTING AND PUBLISHING
I4751	NEWSPAPER PRINTING AND PUBLISHING
I47510	NEWSPAPER PRINTING AND PUBLISHING
I4752	PERIODICAL PRINTING AND PUBLISHING
I47520	PERIODICAL PRINTING AND PUBLISHING
I47521	TRADE JOURNAL PUBLISHING
I4752105	TRADE JOURNAL PUBLISHING
I4753	BOOK PRINTING AND PUBLISHING
I47530	BOOK PRINTING AND PUBLISHING
I48	RUBBER AND PLASTICS PROCESSING
I480	RUBBER AND PLASTICS PROCESSING
I4800	RUBBER AND PLASTICS PROCESSING
I48000	RUBBER AND PLASTICS PROCESSING
I481	RUBBER PRODUCTS
I4810	RUBBER PRODUCTS
I48100	RUBBER PRODUCTS
I4811	TYRE MANUFACTURING
I48110	TYRE MANUFACTURING
I483	PLASTICS PRODUCTS
I4830	PLASTICS PRODUCTS

CODE	DESCRIPTION
I48300	PLASTICS PRODUCTS
I49	PROCESSING INDUSTRIES
I491	JEWELLERY
I4910	JEWELLERY
I49100	JEWELLERY
I492	MUSICAL INSTUMENTS
I4920	MUSICAL INSTUMENTS
I49200	MUSICAL INSTUMENTS
I493	PHOTOGRAPHIC PROCESSING
I4930	PHOTOGRAPHIC PROCESSING
I49300	PHOTOGRAPHIC PROCESSING
I494	PROCESSING INDUSTRIES
I4941	TOYS AND GAMES
I49410	TOYS AND GAMES
I4942	SPORTING EQUIPMENT
I49420	SPORTING EQUIPMENT
I495	STATIONERY, NOT PAPER
I4954	STATIONERY, NOT PAPER
I49540	STATIONERY, NOT PAPER
I5	CONSTRUCTION
I50	CONSTRUCTION
I500	GENERAL CONSTRUCTION AND DEMOLITION
I5000	GENERAL CONSTRUCTION AND DEMOLITION
I50000	GENERAL CONSTRUCTION AND DEMOLITION
I501	CONSTRUCTION OF BUILDINGS
I5010	CONSTRUCTION OF BUILDINGS
I50100	CONSTRUCTION OF BUILDINGS
I5010022	RESIDENTIAL CONSTRUCTION
I5010023	OFFICE CONSTRUCTION
I5010024	HOSPITAL CONSTRUCTION
I5010025	FACTORY AND INDUSTRIAL CONSTRUCTION
I5010027	RETAIL CONSTRUCTION
I5010028	EDUCATIONAL FACILITY CONSTRUCTION
I5010029	BUILDING REFURBISHMENT
I5010031	LEISURE AND ENTERTAINMENTS CONSTRUCTION
I502	CIVIL ENGINEERING
I5020	CIVIL ENGINEERING
I50200	CIVIL ENGINEERING
I5020002	AIRPORT CONSTRUCTION
I5020006	BRIDGE BUILDING
I5020008	CAR PARK CONSTRUCTION
I5020011	DAM CONSTRUCTION
I5020017	HARBOUR CONSTRUCTION
I5020022	OIL PRODUCTION PLATFORM CONSTRUCTION
I5020028	RAILWAY CONSTRUCTION
I502003	RESERVOIR CONSTRUCTION
I5020030	RESERVOIR CONSTRUCTION
I5020032	SEWERAGE CONSTRUCTION
I5020039	TUNNEL CONSTRUCTION
I5020041	ROAD CONSTRUCTION
I5020043	PIPELINE LAYING
I5020044	POWER STATION CONSTRUCTION
I5020045	WATER TREATMENT PLANT CONSTRUCTION
I5020047	LAND RECLAMATION
I502005	SEA DEFENCE CONSTRUCTION
I5020050	SEA DEFENCE CONSTRUCTION
I5020051	CANAL OR WATERWAY CONSTRUCTION
I503	BUILDING INSTALLATIONS
I5030	BUILDING INSTALLATIONS
I50300	BUILDING INSTALLATIONS
I504	BUILDING COMPLETION
I5040	BUILDING COMPLETION
I50400	BUILDING COMPLETION
I6	DISTRIBUTION, HOTELS AND CATERING
I60	DISTRIBUTION, HOTELS AND CATERING
I600	DISTRIBUTION, HOTELS AND CATERING
I6000	DISTRIBUTION, HOTELS AND CATERING
I60000	DISTRIBUTION, HOTELS AND CATERING
I61	WHOLESALE DISTRIBUTION
I610	WHOLESALE DISTRIBUTION
I6100	WHOLESALE DISTRIBUTION
I61000	WHOLESALE DISTRIBUTION
I63	PRODUCT TRADING AND BROKING
I630	PRODUCT TRADING AND BROKING
I6300	PRODUCT TRADING AND BROKING
I63000	PRODUCT TRADING AND BROKING
I64	RETAIL DISTRIBUTION
I640	RETAIL DISTRIBUTION
I6400	RETAIL DISTRIBUTION
I64000	RETAIL DISTRIBUTION
I641	FOOD RETAILING
I6410	FOOD RETAILING
I64100	FOOD RETAILING
I642	CIGARETTE, NEWSPAPER AND LIQUOR STORES
I6420	CIGARETTE, NEWSPAPER AND LIQUOR STORES
I64200	CIGARETTE, NEWSPAPER AND LIQUOR STORES
I643	CHEMISTS AND DRUG STORES
I6430	CHEMISTS AND DRUG STORES
I64300	CHEMISTS AND DRUG STORES
I645	CLOTHING STORES
I6450	CLOTHING STORES
I64500	CLOTHING STORES
I646	SHOE STORES
I6460	SHOE STORES
I64600	SHOE STORES
I647	HOUSEHOLD TEXTILES
I6470	HOUSEHOLD TEXTILES
I64700	HOUSEHOLD TEXTILES
I648	HOUSEHOLD GOODS AND HARDWARE
I6480	HOUSEHOLD GOODS AND HARDWARE
I64800	HOUSEHOLD GOODS AND HARDWARE
I65	RETAIL DISTRIBUTION
I650	RETAIL DISTRIBUTION
I6500	RETAIL DISTRIBUTION
I65000	RETAIL DISTRIBUTION
I651	AUTOMOBILE DEALERS
I6510	AUTOMOBILE DEALERS
I65100	AUTOMOBILE DEALERS
I652	FILLING STATIONS
I6520	FILLING STATIONS
I65200	FILLING STATIONS
I653	BOOKS, STATIONERY RETAILING
I6530	BOOKS, STATIONERY RETAILING
I65300	BOOKS, STATIONERY RETAILING
I654	SPECIALIST STORES
I6540	SPECIALIST STORES
I65400	SPECIALIST STORES
I6540005	OPTICIANS
I6540011	GARDEN CENTRES
I654003	TOYS AND GAMES RETAILING
I6540030	TOYS AND GAMES RETAILING
I656	MIXED RETAIL
I6560	MIXED RETAIL
I65600	MIXED RETAIL
I6560002	DEPARTMENT STORE
I6560003	MAIL ORDER
I6560011	OUT OF TOWN RETAILING
I66	HOTELS AND CATERING
I660	HOTELS AND CATERING
I6600	HOTELS AND CATERING
I66000	HOTELS AND CATERING
I661	RESTAURANTS, CAFES AND FAST FOOD
I6610	RESTAURANTS, CAFES AND FAST FOOD
I66100	RESTAURANTS, CAFES AND FAST FOOD
I662	BARS AND PUBLIC HOUSES
I6620	BARS AND PUBLIC HOUSES
I66200	BARS AND PUBLIC HOUSES
I665	HOTELS AND ACCOMMODATION
I6650	HOTELS AND ACCOMMODATION
I66500	HOTELS AND ACCOMMODATION
I67	REPAIR OF VEHICLES AND CONSUMER GOODS
I670	REPAIR OF VEHICLES AND CONSUMER GOODS

CODE	DESCRIPTION
I6700	REPAIR OF VEHICLES AND CONSUMER GOODS
I67000	REPAIR OF VEHICLES AND CONSUMER GOODS
I7	TRANSPORT AND COMMUNICATION
I70	TRANSPORT AND COMMUNICATION
I700	TRANSPORT AND COMMUNICATION
I7000	TRANSPORT AND COMMUNICATION
I70000	TRANSPORT AND COMMUNICATION
I71	RAILWAYS
I710	RAILWAYS
I7100	RAILWAYS
I71000	RAILWAYS
I72	TRANSPORT AND COMMUNICATION
I721	TRANSPORT AND COMMUNICATION
I7210	TRANSPORT AND COMMUNICATION
I72101	URBAN MASS TRANSIT SYSTEMS
I72102	BUS AND COACH SERVICES
I722	TAXI SERVICES
I7220	TAXI SERVICES
I72200	TAXI SERVICES
I723	ROAD HAULAGE
I7230	ROAD HAULAGE
I72300	ROAD HAULAGE
I726	INLAND WATER TRANSPORT
I7260	INLAND WATER TRANSPORT
I72603	INLAND WATER TRANSPORT
I74	SEA TRANSPORT
I740	SEA TRANSPORT
I7400	SEA TRANSPORT
I74000	SEA TRANSPORT
I75	AIR TRANSPORT
I750	AIR TRANSPORT
I7500	AIR TRANSPORT
I75000	AIR TRANSPORT
I751	SPACE TRANSPORT
I7510	SPACE TRANSPORT
I75100	SPACE TRANSPORT
I76	TRANSPORT AND COMMUNICATION
I763	PORTS AND SHIPPING SUPPORT SERVICES
I7630	PORTS AND SHIPPING SUPPORT SERVICES
I76300	PORTS AND SHIPPING SUPPORT SERVICES
I764	AIRPORTS AND AIR SUPPORT SERVICES
I7640	AIRPORTS AND AIR SUPPORT SERVICES
I76400	AIRPORTS AND AIR SUPPORT SERVICES
I77	TRANSPORT AND COMMUNICATION
I770	TRANSPORT AND COMMUNICATION
I7700	TRANSPORT AND COMMUNICATION
I77001	TRAVEL AGENTS
I77002	FREIGHT FORWARDING
I77003	STORAGE AND WAREHOUSING
I79	TRANSPORT AND COMMUNICATION
I790	TRANSPORT AND COMMUNICATION
I7901	POSTAL SERVICES
I79010	POSTAL SERVICES
I7902	TELECOMMUNICATIONS
I79020	TELECOMMUNICATIONS
I8	FINANCIAL AND BUSINESS SERVICES
I80	FINANCIAL AND BUSINESS SERVICES
I800	FINANCIAL AND BUSINESS SERVICES
I8000	FINANCIAL AND BUSINESS SERVICES
I80000	FINANCIAL AND BUSINESS SERVICES
I81	BANKING AND FINANCIAL SERVICES
I810	BANKING AND FINANCIAL SERVICES
I8100	BANKING AND FINANCIAL SERVICES
I81000	BANKING AND FINANCIAL SERVICES
I814	BANKING SERVICES
I8140	BANKING SERVICES
I81400	BANKING SERVICES
I81401	CENTRAL BANKING AUTHORITIES
I81402	COMMERCIAL BANKING
I81403	SAVINGS BANKS
I815	BANKING AND FINANCIAL SERVICES
I8150	BANKING AND FINANCIAL SERVICES
I81501	NON-BANK CREDIT INSTITUTIONS
I8150103	MORTGAGE INSTITUTIONS
I8150106	DEVELOPMENT BANKS OR FUNDS
I8150108	CREDIT CARD ISSUERS
I815011	EXPORT-IMPORT FINANCING
I8150110	EXPORT-IMPORT FINANCING
I81502	TRUSTS, SECURITIES INVESTMENT
I8150203	VENTURE CAPITAL
I8150206	INVESTMENT TRUSTS
I8150211	UNIT TRUSTS AND MUTUAL FUNDS
I8150214	PENSION FUNDS
I8150216	PERSONAL EQUITY PLANS
I82	INSURANCE
I820	INSURANCE
I8200	INSURANCE
I82000	INSURANCE
I82001	COMPOSITE INSURANCE
I82002	LIFE INSURANCE
I82003	NON-LIFE INSURANCE
I8200316	MOTOR INSURANCE
I8200318	REINSURANCE
I83	BUSINESS SERVICES
I83000	TEMPORARY
I831	FINANCIAL SERVICES
I8310	FINANCIAL SERVICES
I83100	FINANCIAL SERVICES
I832	INSURANCE BROKERS AND AGENTS
I8320	INSURANCE BROKERS AND AGENTS
I83200	INSURANCE BROKERS AND AGENTS
I834	REAL ESTATE AGENTS
I8340	REAL ESTATE AGENTS
I83400	REAL ESTATE AGENTS
I835	LEGAL SERVICES
I8350	LEGAL SERVICES
I83500	LEGAL SERVICES
I836	ACCOUNTANCY AND AUDITING
I8360	ACCOUNTANCY AND AUDITING
I83600	ACCOUNTANCY AND AUDITING
I837	ARCHITECTS, SURVEYORS, ETC
I8370	ARCHITECTS, SURVEYORS, ETC
I83700	ARCHITECTS, SURVEYORS, ETC
I838	ADVERTISING AGENCIES
I8380	ADVERTISING AGENCIES
I83800	ADVERTISING AGENCIES
I839	BUSINESS SERVICES
I83900	TEMPORARY
I8394	COMPUTER SERVICES
I83940	COMPUTER SERVICES
I8394007	COMPUTER MAINTENANCE AND SUPPORT
I8395	BUSINESS SERVICES
I83951	MANAGEMENT CONSULTANTS
I83952	MARKET RESEARCH AND PUBLIC RELATIONS
I8395205	PUBLIC RELATIONS CONSULTANTS
I83953	DOCUMENT COPYING AND DUPLICATING
I83954	BUSINESS AND MEDIA SERVICES
I8395416	FINANCIAL AND CREDIT REPORTING
I8395419	FREE-LANCE JOURNALISTS
I8395448	SALES PROMOTION
I8395449	DIRECT MARKETING
I8395451	MEDIA PLANNING
I8396	DIVERSIFIED HOLDING COMPANIES
I83960	DIVERSIFIED HOLDING COMPANIES
I84	RENTING AND LEASING EQUIPMENT
I840	RENTING AND LEASING EQUIPMENT
I8400	RENTING AND LEASING EQUIPMENT
I84000	RENTING AND LEASING EQUIPMENT
I841	AGRICULTURAL EQUIPMENT HIRE
I8410	AGRICULTURAL EQUIPMENT HIRE
I84100	AGRICULTURAL EQUIPMENT HIRE
I842	CONSTRUCTION EQUIPMENT HIRE
I8420	CONSTRUCTION EQUIPMENT HIRE
I84200	CONSTRUCTION EQUIPMENT HIRE
I843	OFFICE EQUIPMENT HIRE
I8430	OFFICE EQUIPMENT HIRE
I84300	OFFICE EQUIPMENT HIRE

CODE	DESCRIPTION
I846	CONSUMER GOODS HIRE
I8460	CONSUMER GOODS HIRE
I84600	CONSUMER GOODS HIRE
I848	RENTING AND LEASING EQUIPMENT
I8480	RENTING AND LEASING EQUIPMENT
I84800	TEMPORARY
I84801	AUTOMOBILE HIRE
I84802	LIGHT COMMERCIAL VEHICLE HIRE
I84803	COMMERCIAL VEHICLE HIRE
I84804	OTHER TRANSPORT EQUIPMENT HIRE
I8480401	AIRCRAFT HIRING AND LEASING
I848041	SHIP HIRING AND LEASING
I8480410	SHIP HIRING AND LEASING
I85	REAL ESTATE DEALING
I850	REAL ESTATE DEALING
I8500	REAL ESTATE DEALING
I85000	REAL ESTATE DEALING
I8500005	INDUSTRIAL REAL ESTATE DEALING
I8500011	RESIDENTIAL REAL ESTATE DEALING
I8500021	COMMERCIAL REAL ESTATE DEALING
I8500029	RETAIL REAL ESTATE DEALING
I8500031	RESIDENTIAL REAL ESTATE BUYING, SELLING
I9	SERVICES AND ENTERTAINMENT
I90	SERVICES AND ENTERTAINMENT
I900	SERVICES AND ENTERTAINMENT
I9000	SERVICES AND ENTERTAINMENT
I90000	SERVICES AND ENTERTAINMENT
I92	SERVICES AND ENTERTAINMENT
I921	SERVICES AND ENTERTAINMENT
I92100	TEMPORARY
I9211	REFUSE DISPOSAL
I92110	REFUSE DISPOSAL
I9212	SEWAGE DISPOSAL
I92120	SEWAGE DISPOSAL
I923	CLEANING SERVICES
I9230	CLEANING SERVICES
I92300	CLEANING SERVICES
I95	SERVICES AND ENTERTAINMENT
I951	HOSPITALS AND HEALTHCARE
I9510	HOSPITALS AND HEALTHCARE
I95100	HOSPITALS AND HEALTHCARE
I97	SERVICES AND ENTERTAINMENT
I971	FILM PRODUCTION AND DISTRIBUTION
I9710	FILM PRODUCTION AND DISTRIBUTION
I97100	FILM PRODUCTION AND DISTRIBUTION
I974	TELEVISION AND RADIO
I9740	TELEVISION AND RADIO
I97400	TELEVISION AND RADIO
I9741	TELEVISION AND RADIO
I97411	TELEVISION AND RADIO
I9741102	TELEVISION BROADCASTING
I9741105	RADIO BROADCASTING
I9741109	CABLE TELEVISION
I974111	SATELLITE TELEVISION
I9741110	SATELLITE TELEVISION
I9741112	TELEVISION PROGRAMME PRODUCTION
I97412	THEATRES AND LIVE ENTERTAINMENT VENUES
I979	SERVICES AND ENTERTAINMENT
I9791	SERVICES AND ENTERTAINMENT
I97911	SPORTING FACILITIES AND VENUES
I97912	BETTING AND GAMBLING
I98	LAUNDRIES AND DRY CLEANERS
I981	LAUNDRIES AND DRY CLEANERS
I9810	LAUNDRIES AND DRY CLEANERS
I98100	LAUNDRIES AND DRY CLEANERS
I9999	DUMMY CODE
I99999	DUMMY CODE

APPENDIX I: Stop words

accordingly	because	enough	its
again	become	especially	itself
against	becomes	etc	keep
almost	been	every	kept
already	before	for	kg
also	being	from	km
although	between	gave	largely
always	biol	gets	mainly
among	both	give	most
an	briefly	given	mostly
and	by	giving	much
another	came	got	mug
any	certain	hardly	must
anyone	certainly	how	nearly
apparently	chem	however	necessarily
are	different	immediately	neither
arise	during	importance	next
aside	each	important	none
at	either	into	normally
away	else	it	nos

now
obtain
obtained
of
often
on
other
ought
our
owing
particularly
perhaps
please
possible
possibly
potentially
predominantly
previously
primarily
probably
prompt
promptly
quickly
quite
rather
readily
really
recently
refs
regardless
relatively
respectively
said
same
seem
several
should
significantly
similar
similarly
since
slightly
so
some
sometime
somewhat
soon
specifically
strongly
substantially
successfully
such
sufficiently
than
that
the
their
theirs
them
then
there
therefore
these
they
this
those
though
through
throughout
to
too
toward
until
upon
usefully
usefulness
various
was
were
what
when
where
whether
which
while
who
whose
why
widely
with
within
yet

APPENDIX J: Confusion Matrix (Percent)

Confusion Matrix (Percent)

A confusion matrix in percent shows the percentage of predicted classes for each actual class. Take, for example, the 2 by 2 matrix for the training set of Dataset 1. The first row indicates that of all the actual class 1 instances 95,79% was classified class 1 and 4.21% was classified -1. The second row show that of all actual class -1 instances 94.12% were classified -1 and 5.88% were classified class 1. This type of matrix is slightly different from what is commonly known as a confusion matrix, whereby the actual numbers of instances are given.

Polynomial Kernel

	Dataset 1			
	Predicted Class			
	Train		Test	
Actual Class	1	-1	1	-1
1	95,79%	4,21%	59,89%	40,11%
-1	5,88%	94,12%	46,66%	53,34%

	Dataset 2			
	Predicted Class			
	Train		Test	
Actual Class	1	-1	1	-1
1	98,13%	1,87%	71,19%	28,81%
-1	6,64%	93,36%	49,27%	50,73%

	Dataset 3			
	Predicted Class			
	Train		Test	
Actual Class	1	-1	1	-1
1	97,10%	2,90%	64,06%	35,94%
-1	7,86%	92,14%	51,98%	48,02%

	Dataset 4			
	Predicted Class			
	Train		Test	
Actual Class	1	-1	1	-1
1	98,39%	1,61%	66,42%	33,58%
-1	6,26%	93,74%	49,41%	50,59%

	Dataset 5			
	Predicted Class			
	Train		Test	
Actual Class	1	-1	1	-1
1	53,94%	46,06%	14,40%	85,60%
-1	0,38%	99,62%	5,76%	94,24%

	Dataset 6			
	Predicted Class			
	Train		Test	
Actual Class	1	-1	1	-1
1	98,16%	1,84%	72,62%	27,38%
-1	6,45%	93,55%	53,06%	46,94%

	Dataset 7			
	Predicted Class			
	Train		Test	
Actual Class	1	-1	1	-1
1	98,28%	1,72%	71,77%	28,23%
-1	11,91%	88,09%	63,20%	36,80%

	Dataset 8			
	Predicted Class			
	Train		Test	
Actual Class	1	-1	1	-1
1	98,33%	1,67%	73,88%	26,12%
-1	5,84%	94,16%	48,82%	51,18%

	Dataset 9			
	Predicted Class			
	Train		Test	
Actual Class	1	-1	1	-1
1	88,01%	11,99%	54,02%	45,98%
-1	1,63%	98,37%	18,77%	81,23%

	Dataset 10			
	Predicted Class			
	Train		Test	
Actual Class	1	-1	1	-1
1	99,20%	0,80%	86,79%	13,21%
-1	11,63%	88,37%	44,48%	55,52%

Rbf Kernel

	Dataset 1			
	Predicted Class			
	Train		Test	
Actual Class	1	-1	1	-1
1	98,25%	1,75%	63,44%	36,56%
-1	2,65%	97,35%	52,49%	47,51%

	Dataset 2			
	Predicted Class			
	Train		Test	
Actual Class	1	-1	1	-1
1	99,12%	0,88%	85,48%	14,52%
-1	2,07%	97,93%	70,85%	29,15%

	Dataset 3			
	Predicted Class			
	Train		Test	
Actual Class	1	-1	1	-1
1	98,15%	1,85%	78,97%	21,03%
-1	2,56%	97,44%	68,89%	31,11%

	Dataset 4			
	Predicted Class			
	Train		Test	
Actual Class	1	-1	1	-1
1	99,22%	0,78%	79,85%	20,15%
-1	1,65%	98,35%	72,06%	27,94%

	Dataset 5			
	Predicted Class			
	Train		Test	
Actual Class	1	-1	1	-1
1	95,00%	5,00%	12,62%	87,38%
-1	0,55%	99,45%	3,55%	96,45%

	Dataset 6			
	Predicted Class			
	Train		Test	
Actual Class	1	-1	1	-1
1	99,17%	0,83%	86,19%	13,81%
-1	1,85%	98,15%	74,93%	25,07%

	Dataset 7			
	Predicted Class			
	Train		Test	
Actual Class	1	-1	1	-1
1	98,70%	1,30%	85,19%	14,81%
-1	3,38%	96,62%	76,16%	23,84%

	Dataset 8			
	Predicted Class			
	Train		Test	
Actual Class	1	-1	1	-1
1	99,25%	0,75%	89,30%	10,70%
-1	1,98%	98,02%	72,94%	27,06%

	Dataset 9			
	Predicted Class			
	Train		Test	
Actual Class	1	-1	1	-1
1	97,64%	2,36%	35,96%	64,04%
-1	0,80%	99,20%	10,27%	89,73%

	Dataset 10			
	Predicted Class			
	Train		Test	
Actual Class	1	-1	1	-1
1	99,63%	0,37%	93,76%	6,24%
-1	1,34%	98,66%	68,51%	31,49%

APPENDIX K: Software Used

Raw Dataset		
Purpose	**Filename**	**Software**
Extracting Market Data - S&P 500 - Exxon Mobil Corp. Prices Trading Volume	 marketindex.mat workdata1.mat	 MATLAB 7.0 R14 MATLAB 7.0 R14
DataStream Data - Exxon Mobil Corp Prices	 TAN.xls	 Micosoft Excel 2003
News in NewsML format	*.xml	Internet Explorer

Processing		
Purpose	**Filename**	**Software**
Creating Market Labels	labels.xls	Micosoft Excel 2003
List of Labels	labels.txt	Notepad
List of Trading Day Dates	dates.csv	MATLAB 7.0 R14
Unpack NewsML files from Reuters CD	UnpackNML.java	Java 1.4.2
Select NewsML by Industry	IndustrySelect.java	Java 1.4.2

Processing		
Purpose	**Filename**	**Software**
Select only trading days	SelectTradingDays.java	Java 1.4.2
Create Corpus for Term Extraction	CreateCorpus.java	Java 1.4.2
Annotate Corpus with Part-of-Speech tags	Annotator.java	Term Extractor Java 1.4.2
Filter Terms from Annotated Corpus	LinguisticFilter.java	Term Extractor Java 1.4.2
Use list of terms to create Lexicon for Collexis® Engine	builder.exe	Collexis® Thesaurus Builder 5.5
Test the created Lexicon	AbstractionComponentTester.exe	Collexis® Abstraction Component Test 5.0
Rename XML extension to HTML	RenameXML.java	Java 1.4.2
Create list of available NewsML files	Importlist.java	Java 1.4.2
Import NewsML into Collexis® Engine	ImportBulk.exe	Collexis® ImportBulk 5.0
Export Fingerprints	exportCorrectFP.py	Python 2.2.1
Due to error some fingerprints had to be re-exported	exportremainingFP.py	Python 2.2.1
Re-exported fingerprints had to be corrected by Collexis® engine	correctFP.py	Python 2.2.1
Create Fingerprint Dataset by getting concept and rank	CreateSVMData.java	Java 1.4.2
Label the Fingerprint Dataset	LabelSVMData.java	Java 1.4.2
Modelling SVM in Matlab	osu_svm3.00.zip	MATLAB 7.0 R14