2007:071

MASTER'S THESIS

Stock Trend Prediction Using News Articles

A Text Mining Approach

Pegah Falinouss

Luleå University of Technology

Master Thesis, Continuation Courses
Marketing and e-commerce
Department of Business Administration and Social Sciences
Division of Industrial marketing and e-commerce

2007:071 - ISSN: 1653-0187 - ISRN: LTU-PB-EX--07/071--SE

Stock Trend Prediction Using News Articles
A Text Mining Approach

Supervisors:
Dr. Mohammad Sepehri
Dr. Moez Limayem

Prepared by: Pegah Falinouss

Tarbiat Modares University
Faculty of Engineering
Department of Industrial Engineering

Luleå University of Technology
Department of Business Administration and Social Sciences
Division of Industrial Marketing and E-Commerce

MSc PROGRAM IN MARKETING AND ELECTRONIC COMMERCE

Joint 2007

Abstract

Stock market prediction with data mining techniques is one of the most important issues to be investigated. Mining textual documents and time series concurrently, such as predicting the movements of stock prices based on the contents of the news articles, is an emerging topic in data mining and text mining community. Previous researches have shown that there is a strong relationship between the time when the news stories are released and the time when the stock prices fluctuate.

In this thesis, we present a model that predicts the changes of stock trend by analyzing the influence of non-quantifiable information namely the news articles which are rich in information and superior to numeric data. In particular, we investigate the immediate impact of news articles on the time series based on the Efficient Markets Hypothesis. This is a binary classification problem which uses several data mining and text mining techniques.

For making such a prediction model, we use the intraday prices and the time-stamped news articles related to Iran-Khodro Company for the consecutive years of 1383 and 1384. A new statistical based piecewise segmentation algorithm is proposed to identify trends on the time series. The news articles are preprocessed and are labeled either as rise or drop by being aligned back to the segmented trends. A document selection heuristics that is based on the chi-square estimation is used for selecting the positive training documents. The selected news articles are represented using the vector space modeling and tfidf term weighting scheme. Finally, the relationship between the contents of the news stories and trends on the stock prices are learned through support vector machine.

Different experiments are conducted to evaluate various aspects of the proposed model and encouraging results are obtained in all of the experiments. The accuracy of the prediction model is equal to 83% and in comparison with news random labeling with 51% of accuracy; the model has increased the accuracy by 30%. The prediction model predicts 1.6 times better and more correctly than the news random labeling.

Acknowledgment

There are many individuals who contributed to the production of this thesis through their moral and technical support, advice, or participation.

I am indebted to my supervisors **Dr. Mehdi Sepehri** and **Dr. Moez Limayem** for their patience, careful supervision, and encouragement throughout the completion of my thesis project. It has been both a privilege and a pleasure to have experienced the opportunity to be taught by two leading international scholars. I sincerely thank you both for being the sort of supervisors every student needs - astute, supportive, enthusiastic, and inspiring. The ideal role models for a beginning academic and the best possible leading academics to supervise an ambitious enhancement study.

I would like to express my appreciation to **Dr. Babak Teimourpour**, the PhD student in Industrial Engineering in Tarbiat Modares University. He has been of great help, support, and encouragement in accomplishing the research process.

I would also like to express my gratitude to **Tehran Stock Exchange Services Company** for their cooperation in providing the data from their databases.

Finally, I would like to thank my family and friends and especially my husband for his understanding, encouragement, and support over the completion and fulfillment of my research project. I would like to dedicate my thesis to my parents and my husband.

Table of Content

List of Table

List of Figure

Chapter 1

Introduction and Preface

1. Introduction and Preface

The rapid progress in digital data acquisition has led to the fast-growing amount of data stored in databases, data warehouses, or other kinds of data repositories. (Zhou, 2003) Although valuable information may be hiding behind the data, the overwhelming data volume makes it difficult for human beings to extract them without powerful tools. In order to relieve such a data rich but information poor dilemma, during the late 1980s, a new discipline named data mining emerged, which devotes itself to extracting knowledge from huge volumes of data, with the help of the ubiquitous modern computing devices, namely, computer. (Markellos et al., 2003)

1.1 Considerations and Background

Financial time series forecasting has been addressed since the 1980s. The objective is to beat financial markets and win much profit. Until now, financial forecasting is still regarded as one of the most challenging applications of modern time series forecasting. Financial time series have very complex behavior, resulting from a huge number of factors which could be economic, political, or psychological. They are inherently noisy, non-stationary, and deterministically chaotic. (Tay et al., 2003)

Due to the complexity of financial time series, there is some skepticism about the predictability of financial time series. This is reflected in the well-known efficient market hypothesis theory (EMH) introduced by Fama (1970). According to the EMH theory, the current price is the best prediction for the next day, and buy-hold is the best trading strategy. However, there are strong evidences which refuse the efficient market hypothesis. Therefore, the task is not to doubt whether financial time series are predictable, but to discover a good model that is capable of describing the dynamics of financial time series.

The number of proposed methods in financial time series prediction is tremendously large. These methods rely heavily in using structured and numerical databases. In the field of trading, most analysis tools of the stock market still focus on statistical analysis of past price developments. But one of the areas in stock market prediction comes from textual data, based on the assumption that the course of a stock price can be predicted much better by looking at appeared news articles. In stock market, the share prices can be influenced by many factors, ranging from news releases of companies and local politics to news of superpower economy. (Ng and Fu, 2003)

Easy and quick availability to news information was not possible until the beginning of the last decade. In this age of information, news is now easily accessible, as content providers and content locators such as online news services have sprouted on the World Wide Web. Nowadays, there is a large amount of information available in the form of text in diverse environments, the analysis of which can provide many benefits in several areas. (Hariharan, 2004) The continuous availability of more news articles in digital form, the latest developments in Natural Language Processing (NLP) and the availability of faster computers lead to the question how to extract more information out of news articles. (Bunningen, 2004) It seems that there is a need for extending the focus to mining information from unstructured and semi-structured information sources. Hence, there is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of unstructured digital data. These theories and tools are the subject of the emerging field of knowledge discovery in text databases, known as text mining.

Knowledge Discovery in Databases (KDD), also known as *data mining*, focuses on the computerized exploration of large amounts of data and on the discovery of interesting patterns within them. Until recently computer scientists and information system specialists concentrated on the discovery of knowledge from structured, numerical databases and data warehouses. However, a lot of information nowadays is available in the form of text, including documents, news, manuals, email, and etc. The increasing number of textual data has led to knowledge discovery in unstructured (textual databases) data known as text mining or text data mining (Hearst, 1997). Text mining is an emerging technology for analyzing large collections of unstructured documents for the purposes of extracting interesting and non-trivial patterns or knowledge. Text mining has a goal to look for patterns in natural language text and to extract corresponding information. Zorn et al. (1999) regard text mining as a knowledge creation tool which offers powerful possibilities for creating knowledge and relevance out of the massive amounts of unstructured information available on the Internet and corporate intranets.

One of the applications of text mining is discovering and exploiting the relationship between the document text and an external source of information such as time stamped streams of data namely stock market quotes. Predicting the movements of stock prices based on the contents of news articles is one of many applications of text mining techniques. Information about company's report or breaking news stories can dramatically affect the share price of a security. There have been many researches conducted to investigate the influence of news articles on stock market and the reaction of stock market to press releases. Researchers have shown that there is a strong relationship between the time when the news stories are released and the time when the stock prices fluctuate. This made researchers enter to a new area of research, predicting the stock trend movement based on the content of news stories. While there are many promising forecasting methods to predict stock market movements based on numeric time series data, the number of predicting methods concerning the application of text mining techniques using news articles is few. This is because text mining seems to be more complex than data mining as it involves dealing with text data that are inherently unstructured and fuzzy.

1.2 The Importance of Study

Stock markets have been studied over and over again to extract useful patterns and predict their movements. Stock market prediction has always had a certain appeal for researchers and financial investors. The reason is that who can beat the market, can gain excess profit. Financial analysts who invest in stock markets usually are not aware of the stock market behavior. They are facing the problem of stock trading as they do not know which stocks to buy and which to sell in order to gain more profits. If they can predict the future behavior of stock prices, they can act immediately upon it and make profit.

The more accurate the system predicts the stock price movement, the more profit one can gain from the prediction model. Stock price trend forecasting based solely on the technical and fundamental data analysis enjoys great popularity. But numeric time series data only contain the event and not the cause why it happened. Textual data such as news articles have richer information, hence exploiting textual information especially in addition to numeric time series data increases the quality of the input and improved predictions are expected from this kind of input rather than only numerical data.

Without the doubt, human behaviors are always influenced by their environment. One of the most significant impacts that affect the humans' behavior comes from the mass media or to be more specific, from the news articles. On the other hand, the movements of prices in financial markets are the consequences of the actions taken by the investors on how they perceive the events surrounding them as well as the financial markets. As news articles will influence the humans' decision and humans' decision will influence the stock prices, news articles will in turn affect the stock market indirectly.

An increasing amount of crucial and valuable real-time news articles highly related to the financial markets is widely available on the Internet. Extracting valuable information and figuring out the relationship between the extracted information and the financial markets is a critical issue, as it helps financial analyst predict the stock market behavior and gain excess profit. Stock brokers can make their customers more satisfy by offering them the profitable trading rules.

1.3 Problem Statement

Financial analysts who invest in stock markets usually are not aware of the stock market behavior. They are facing the problem of stock trading as they do not know which stocks to buy and which to sell in order to gain more profits. All these users know that the progress of the stock market depends a lot on relevant news and they have to deal daily with vast amount of information. They have to analyze all the news that appears on newspapers, magazines and other textual resources. But analysis of such amount of financial news and articles in order to extract useful knowledge exceeds human capabilities. Text mining techniques can help them automatically extracting the useful knowledge out of textual resources.

Considering the assumption that news articles might give much better predictions of the stock market than analysis of past price developments, and in contrast to the traditional time series analysis, where predictions are made based solely on the technical and fundamental data, we want to investigate the effects of textual information in predicting the financial markets. We would develop a system which is able to use text mining techniques to model the reaction of the stock market to news articles and predict their reactions. By doing so, the investors are able to foresee the future behavior of their stocks when relevant news are released and act immediately upon them.

As input we use real-time news articles and intra-day stock prices of some companies in Tehran Stock Exchange. From these a correlation between certain features found in these articles and changes in stock prices would be made and the predictive model is learned through an appropriate text classifier. Then we feed the system with new news articles and hope that the features found in these articles will cause the same reaction as in the past. Hence the prediction model will notify the up or down of the stock price movement when upcoming news is released and investors can act upon it in order to gain more profit. To find the relationship between stock price movement and the features in news articles, appropriate data and text mining techniques would be applied and different programming languages is used to implement the different data and text mining techniques.

1.4 Research Objective

The financial market is a complex, evolutionary, and non-linear dynamical system. The field of financial forecasting is characterized by data intensity, noise, non-stationary, unstructured nature, high degree of uncertainty, and hidden relationships. Many factors interact in finance including political events, general economic conditions, and traders' expectations. Therefore, predicting price movement in financial markets is quite difficult.

The main objective of this research is to answer the question of how to predict the reaction of stock market to news article, which are rich in valuable information and are more superior to numeric data. To investigate the influence of news articles on stock price movement, different data and text mining techniques are implemented to make the prediction model. With the application of these techniques the relationship between the news features and stock prices are found and a prediction system would be learned using text classifier. Feeding the system with upcoming news, it forecasts the stock price trend. In order to make the prediction model, an extensive programming is required to implement the data and text mining algorithms. All the programming are then combined together to make the whole prediction package. This can be very beneficial for investors, financial analysts, and users of financial news. With such a model they can foresee the movement of stock prices and can act properly in their trading. Moreover this research aims to show that how much valuable information exists in textual databases which with the help of text mining techniques can be extracted and used for various purposes. The overall purpose of study can be summarized in the following research questions:

- How to predict the reaction of stock price trend using textual financial news?
- How data and text mining techniques help to generate this predictive model?

In order to investigate the impact of news on a stock trend movement, we have to make a prediction model. To make the prediction model, we have to use different data and text mining techniques and in order to implement these techniques; we have to use different programming languages. Different steps in the research process are programmed and coded and are combined together to make the prediction model.

1.5 Tehran Stock Exchange (TSE)

As was mentioned in previous section, the objective of this study is to predict the movement of stock price trend based on financial and political news articles. The stocks whose price movements are going to be predicted are those traded in Tehran Stock Market.

The Tehran Stock Exchange opened in April 1968. Initially only Government bonds and certain state-backed certificates were traded in the market. During 1970's the demand for capital boosted the demand for stocks. At the same time institutional changes such as transferring the of shares of public companies and large private firms owned by families to the employees and the private sector led to the expansion of the stock market activity. The restructuring of the economy following the Islamic Revolution expanded public sector control over the economy and reduced the need for private capital. As a result of these events, Tehran Stock Exchange started a period of standstill. This stop came to an end in 1989 with the revitalization of the private sector through privatization of state-owned enterprises and promotion of private sector economic activity based on the First Five-year Development Plan of the country. Since then the Stock Exchange has expanded continuously. Trading in TSE is based on orders sent by the brokers. Presently, TSE trades mainly in securities offered by listed companies available at www.tse.ir. TSE Services Company (TSESC) is in charge of computerized site and supplies computer Services. (Tehran Stock Exchange, 2005)

1.6 Research Orientation

As an introduction to the domain and to place our project in perspective, first we discuss the related work in the area of stock trend prediction with the application of text mining techniques in Chapter 2. In Chapter 3, an overview of time-series preprocessing would be explained. Chapter 4 comprehensively addresses text categorization task and reviews feature selection criteria. Chapter 5 illustrates the overall methodology of our study and Chapter 6 specifies the results and analysis of the proposed model. Conclusions and future works are brought in Chapter 7.

Chapter 2

Literature Review

2. Literature Review

As explained in Chapter 1, the purpose of this study is the prediction of stock trend movement using financial and political news stories. In this chapter the major groundwork and preliminaries related to the subject of the study, is going to be reviewed. We find it necessary to provide a general overview of text mining and stock market movement beforehand.

2.1 Knowledge Discovery in Databases (KDD)

The phrase *"knowledge discovery in databases"* was invented at the first KDD workshop in 1989 (Frawley et al., 1991, 1992) to emphasize that knowledge is the end product of a data-driven discovery. **Knowledge discovery** is defined as the non-trivial extraction of implicit, unknown, and potentially useful information from data. (Frawley et al., 1991, 1992; Fayyad et al., 1996a, 1996b)

Across a wide variety of fields, data are being collected and accumulated at a dramatic pace. There is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data. These theories and tools are the subject of the emerging

field of knowledge discovery in databases. KDD is the intersection of research fields such as machine learning, pattern recognition, databases, statistics, artificial intelligence (AI), knowledge acquisition for expert systems, data visualization, and high-performance computing. (Fayyad et al., 1996b)

Data mining is a step in the KDD process that consists of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns (or models) over the data. (Fayyad et al., 1996a; Parker et al., 1998)

Data mining is a concept that has been establishing itself since the late 1980's. It covers a range of techniques for the efficient discovery of this valuable, non-obvious information from such large collections of data. Essentially, data mining is concerned with the analysis of data and the use of software techniques to find patterns and regularities in datasets. (Parker et al., 1998)

KDD comprises many steps, including **data selection**, **preprocessing**, **transformation**, **data mining**, and **evaluation**, all repeated in multiple iterations. (Figure 2.1) The detailed review of these steps is provided in Chapter 4.

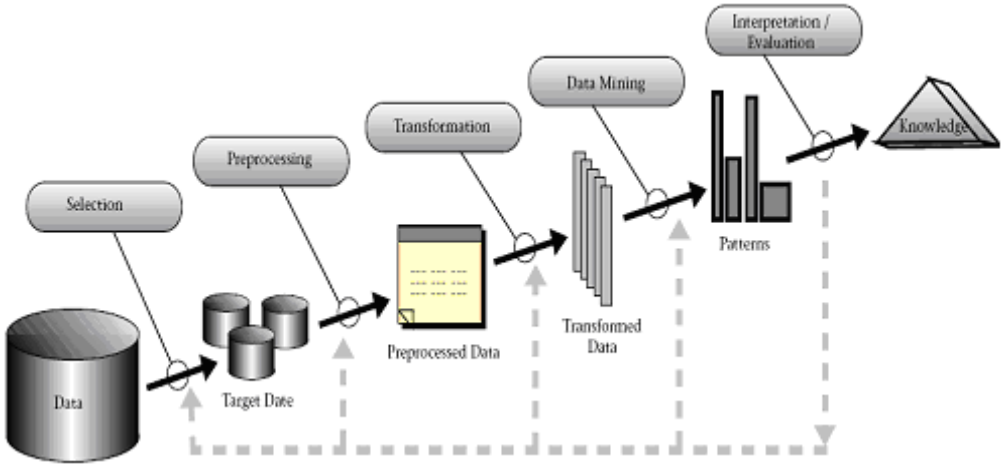


Figure 2.1: An Overview of Steps in KDD Process

Source: Fayyad et al., 1996b

2.1.1 Knowledge Discovery in Text (KDT)

Karanikas and Theodoulidis, (2002) use the term **KDT** to indicate the overall process of turning unstructured textual data into high level information and knowledge, while the term **Text Mining** is used for the step of the KDT process that deals with the extraction of patterns from textual data. By extending the definition of KDD given by Fayyad et al. (1996b), the following simple definition is given: Knowledge Discovery in Text (KDT) is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in **unstructured textual** data.

Text Mining (TM) also known as text data mining (Hearst, 1997) is a step in the KDT process consisting of particular data mining and Natural Language Processing (NLP) algorithms that produces a particular enumeration of patterns over a set of unstructured textual data. (Karanikas and Theodoulidis, 2002) There are various definitions and terminologies for text mining provided by different researchers such as Sullivan (2000), Hearst (1999), Biggs (2000), Albrecht and Merkl (1998) and Zorn et al. (1999)

KDT is a multi-step process, which includes all the tasks from gathering of documents to the visualization and evaluation of the extracted information. The steps are discussed in details in Chapter 4. (Figure 2.2)

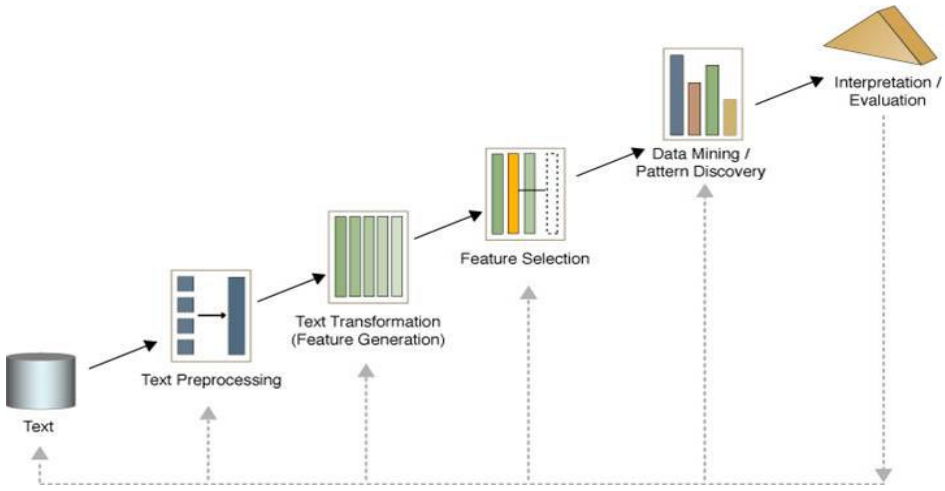


Figure 2.2: KDT Process

Source: Even-Zohar, 2002

2.1.2 Data Mining Vs. Text Mining

Until recently computer scientists and information system specialists concentrated on the discovery of knowledge from structured, numerical databases and data warehouses. However, much, if not the majority, of available business data are captured in text files that are not overtly structured. (Kroeze et al., 2003)

According to Wen (2001) text mining is analogous to data mining in that it uncovers relationships in information. However, unlike data mining, text mining works on information stored in a collection of text documents. Hearst (2003) states that "The difference between regular data mining and text mining is that in text mining the patterns are extracted from unstructured text rather than from structured databases of facts". Dorre et al. (1999) declare that text mining applies the same analytical functions of data mining to the domain of textual information, relying on sophisticated text analysis techniques that distill information from free-text documents.

In conclusion, text mining is similar to data mining in terms of dealing with large volumes of data, and both fall into the information discovery area. The difference between them is that text mining is looking for patterns in unstructured text data, whereas data mining extracts patterns from structured data. Data mining is more mature, while text mining is still in its infancy. (Wen, 2001) Text mining seems to be more complex than data mining as it involves dealing with text data that are inherently unstructured and fuzzy.

2.1.3 The Burgeoning Importance of Text Mining

The area of Knowledge Discovery in Text (KDT) and Text Mining (TM) is growing rapidly mainly because of the strong need for analyzing the vast amount of textual data that reside on internal file systems and the Web. (Karanikas and Theodoulidis, 2002)

In today's information age, we have witnessed and experienced an ever increasing flood of information. The Internet makes available a tremendous amount of information,

on an amazing variety of topics that has been generated for human consumption. Unfortunately, the hundreds of millions of pages of information make it difficult to find information of interest to specific users or useful for particular purposes. The amount of text is simply too large to read and analyze easily. Furthermore, it changes constantly, and requires ongoing review and analysis if one wants to keep abreast of up-to-date information. Working in this ever-expanding sea of text becomes extremely difficult. (Wen, 2001)

As stated by Grobelnik et al. (2000) with the emergence of the World Wide Web, there is a need for extending the focus to mining information from unstructured and semi-structured information sources such as on-line news feeds, corporate archives, research papers, financial reports, medical records, e-mail messages, and etc.

While the amount of textual data available to us is constantly increasing, our ability to understand and process this information remains constant. According to Tan (1999), approximately, 80% of information of an organization is stored in unstructured textual forms such as reports, emails, etc. (Figure 2.3) The need for automated extraction of useful knowledge from huge amounts of textual data in order to assist human analysis is fully apparent. (Merrill Lynch, 2000 cited by Karanikas and Theodoulidis, 2002)

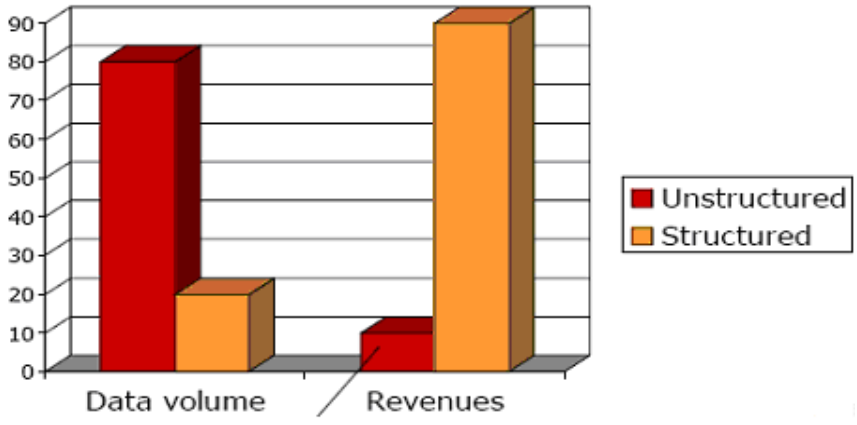


Figure 2.3: Unstructured vs. Structured Data

Source: Raghavan, 2004

2.1.4 Main Text Mining Operations

The main goal of text mining is to enable users extract information from large textual resources. Natural language processing, data mining and machine learning techniques work together to automatically discover patterns from the documents. Most text mining objectives fall under the following categories of operations: Search and Retrieval, Categorization (Supervised Classification), Clustering (Unsupervised Classification), Summarization, Trends Analysis, Associations Analysis, Visualization and etc., The purpose of this study lies on text categorization which is reviewed thoroughly in Chapter 4. For the sake of time and space, we are not discussing other text mining applications.

2.2 Stock Market Movement

Stock markets have been studied over and over again to extract useful patterns and predict their movements. (Hirshleifer and Shumway, 2003) Stock market prediction has always had a certain appeal for researchers. While numerous scientific attempts have been made, no method has been discovered to accurately predict stock price movement. There are various approaches in predicting the movement of stock market and a variety of prediction techniques has been used by stock market analysts. In the following sections, we briefly explain the two most important theories in stock market prediction. Based on these theories two conventional approaches to financial market prediction have emerged: Technical and Fundamental analysis (trading philosophies). The distinction between these two approaches will be also stated.

2.2.1 Theories of Stock Market Prediction

When predicting the future prices of stock market securities, there are two important theories available. The first one is Efficient Market Hypothesis (EMH) introduced by Fama (1964) and the second one is Random Walk Theory. (Malkiel, 1996) The following sections gives the distiction between these two common theories.

2.2.1.1 Efficient Market Hypothesis (EMH)

Fama's contribution in efficient market hypothesis is significant. The Efficient Market Hypothesis (EMH) states that the current market price reflects the assimilation of all the information available. This means that given the information, no prediction of future changes in the price can be made. As new information enters the system the unbalanced state is immediately discovered and quickly eliminated by the correct change in the price. (Fama, 1970) Fama's theory breaks EMH into three forms: Weak, Semi-Strong, and Strong. (Schumaker and Chen, 2006)

In Weak EMH, only past price and historical information is embedded in the current price. This kind of EMH rules out any form of predictions based on the price data only, since the prices follow a randon walk in which successive changes have zero correlation. The Semi-Strong form goes a step further by incorporating all historical and currently public information into the price.This includes additional trading information such as volume data, and fundamental data such as profit prognoses and sales forecast. The Strong form includes historical, public and private information, such as insider information, in the share price.

The weak and semi-strong form of EMH has been fairly supported in a number of research studies. (Low and Webb, 1991; White, 1988). But in recent years many published reports show that Efficent Market Hypothesis is far from correct. Fama (1991) in his article " Efficient Capital Market" states that the efficient market hypothesis surely must be false. The strong form, due to the shortage in data, has been difficult to be tested.

2.2.1.2 Random Walk Theory

A different perspective on prediction comes from Random Walk Theory. (Malkiel 1996) In this theory, stock market prediction is believed to be impossible where prices are determined randomly and outperforming the market is infeasible. Random Walk Theory has similar theoretical underpinning to Semi-String EMH where all public information is assumed to be available to everyone. However, Random Walk Theory declares that even with such information, future prediction is ineffective.

2.2.2 Approaches to Stock Market Prediction

From EMH and Random Walk theories, two distinct trading philosophies have been emerged. These two conventional approaches to financial market prediction are technical analysis and fundamental analysis. In the following sections the distinction between these two approaches will be stated.

2.2.2.1 Technicians Trading Approach

The term technical analysis denotes a basic approach to stock investing where the past prices are studied, using charts as the primary tool. It is based on mining rules and patterns from the past prices of stocks which is called mining of financial time series. The basic principles include concepts such as the trending nature of prices, confirmation and divergence, and the effect of traded volume. Many hundreds of methods for prediction of stock prices have been developed and are still being developed on the grounds of these basic principles. (Hellmstrom and Holmstrom, 1998)

Technical analysis (Pring, 1991) is based on numeric time series data and tries to forecast stock markets using indicators of technical analysis. It is based on the widely accepted hypothesis which says that all reactions of the market to all news are contained in real-time prices of stocks. Because of this, technical analysis ignores news. Its main concern is to identify the existing trends and anticipate the future trends of the stock market from charts. But charts or numeric time series data only contain the event and not the cause why it happened. (Kroha and Baeza-Yates, 2004)

In technical analysis, it is believed that market timing is critical and opportunities can be found through the careful averaging of historical price and volume movements and comparing them against current prices. Technicians utilize charts and modeling techniques to identify trends in price and volume. They rely on historical data in order to predict future outcomes. (Schumaker and Chen, 2006)

There are many promising forecasting methods developed to predict stock market movements from numeric time series. Autoregressive and moving average are some of

the famous stock trend prediction techniques which have dominated the time series prediction for several decays. A thorough survey of the most common technical indicators can be found in the book called “Technical Analysis from A to Z”. (Achelis, 1995)

2.2.2.2 Fundamentalist Trading Approach

Fundamental analysis (Thomsett, 1998) investigates the factors that affect supply and demand. The goal is to gather and interpret this information and act before the information is incorporated in the stock price. The lag time between an event and its resulting market response presents a trading opportunity. Fundamental analysis is based on economic data of companies and tries to forecast markets using economic data that companies have to publish regularly, i.e. annual and quarterly reports, auditor’s reports, balance sheets, income statements, etc. News has an importance for investors using fundamental analysis because news describes factors that may affect supply and demand.

In the fundamentalist trading philosophy, the price of a security can be determined through the nuts and bolts of financial numbers. These numbers are derived from the overall economy, the particular industry’s sector, or most typically, from the company itself. Figures such as inflation, industry return on equity (ROE) and debt levels can all play a part in determining the price of a stock. (Schumaker and Chen, 2006)

One of the areas of limited success in stock market prediction comes from textual data and the use of news articles in price prediction. Information about company’s report or breaking news stories can dramatically affect the share price of a security. There have been many researches conducted to investigate the influence of news articles on stock market and the reaction of stock market to press releases. The overall studies show that stock market reacts to news and the results achieved from previous studies indicate that news articles affect the stock market movement. In the following section, we review some of the researches concerning the influence of new stories on stock prices and volumes traded.

2.2.3 Influence of News Articles on Stock Market

Market and stock exchange news are special messages containing mainly economical and political information. Some of them are carrying information that is important for market prediction. There are various types of financial information sources on the Web which provide electronic versions of their daily issues. All these information sources contain global and regional political and economic news, citations from influential bankers and politicians, as well as recommendations from financial analysts.

Chan et al. (2001) confirm the reaction to news articles. They have shown that economic news always has a positive or negative effect on the number of traded stock. They used salient political and economic news as proxy for public information. They have found that both types of news have impact on measures of trading activity including return volatility, price volatility, number of shares traded, and trading frequency.

Klibanoff et al. (1998) investigate the relationship between closed-end country funds' prices and country-specific salient news. The news that occupies at least two columns wide on *The New York Times* front-page is considered as salient news. They have found that there is a positive relationship between trading volume and salient news. Chan and John-Wei (1996) document that news appearing on the front-page of the *South China Morning Post,* increases the return volatility in the Hong Kong stock market.

Mitchell and Mulherin (1994) use the daily number of headlines reported by Dow Jones as a measure of public information. Using daily data on stock returns and trading volume, they find that market activity is affected by the arrival of news. They report that salient news has a positive impact on absolute price changes.

Berry and Howe (1994), use the number of news released by Reuter's News Service measured in per unit of time as a proxy for public information. In contrast to Mitchell and Mulherin (1994), they look into the impact of news on the intraday market activity. Their results suggest that there is a significant positive relationship between news arrivals and trading volume.

2.3 The Scope of Literature Review

The researches have proven that salient financial and political news affects the stock market and its different attributes including price. This made researchers enter into a new area of research, predicting stock price movement based on news articles. Before evolution of text mining techniques, data mining and statistical techniques were used to forecast the stock market based on only past prices. Their major weakness in that they rely heavily on structural data, which neglects the influence of non-quantifiable information. One can refer to Figure 2.4 for better understanding of what is exactly the scope of this research and what would be the literature review mainly about. As the figure implies, the stock market prediction based on only past prices are out of the scope of this research. The main focus of this research relies on the application of text mining techniques in prediction of stock price movement.

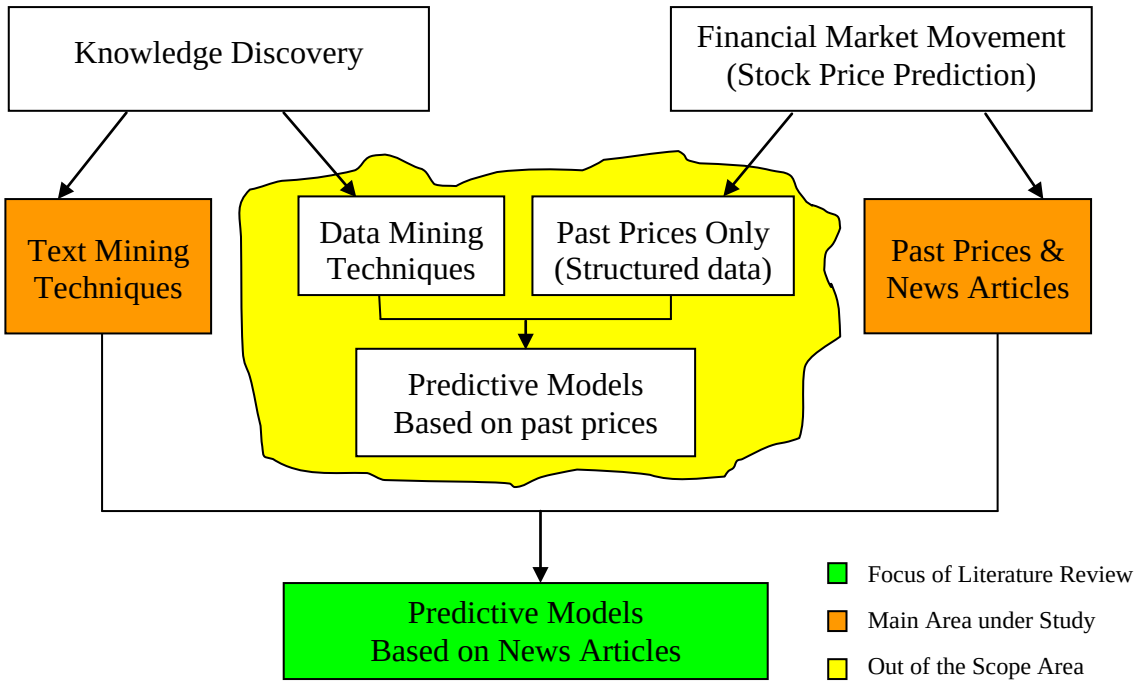


Figure 2.4: The Scope of Literature Review

2.3.1 Text Mining Contribution in Stock Trend Prediction

While there are many articles about data mining techniques in prediction of stock prices, the number of papers concerning the application of text mining in stock market prediction is few. Several papers and publications related to the area of this research have been found and the most important and relevant ones are going to be discussed in the following section. We have provided a list of articles, their authors, and the publication year in Table 2.1. Some PhD and Master's thesis related directly to the scope of this research have been used and reviewed in our study. As the number of scholars in this research area is few, we have prepared a knowledge map introducing the researchers and their contribution to stock trend prediction using news articles. The knowledge map is illustrated in Figure 2.12 at the end of this chapter.

Table 2.1: Articles Related to the Prediction of Stock Market Using News Articles

Articles	Authors
Daily Stock Market Forecast from Textual Web Data	Wuthrich, 1998
Activity Monitoring: Noticing Interesting Changes in Behavior	Fawcett,1999
Electronic Analyst of Stock Behavior (Ænalyst)	Lavrenko,1999
Language Models for Financial News Recommendation	Lavrenko, 2000
Mining of Concurrent Text and Time Series	Lavrenko, 2000
Integrating Genetic Algorithms and Text Learning for Prediction	Sycara et al. 2000
Using News Articles to Predict Stock Price Movements	Gidofalvi, 2001
News Sensitive Stock Trend Prediction	Fung et al. 2002
Stock prediction: Integrating Text Mining Approach Using News	Fung et al. 2003
Forecasting Intraday Stock Price Trends with Text-mining	Mittermayer,2004
Stock Broker P – Sentiment Extraction for the Stock Market	Khare et al, 2004
The Predicting Power of Textual Information on Financial Markets	Fung et al. 2005
Text Mining for Stock Movement Prediction-a Malaysian Approach	Phung, 2005
Textual Analysis of Stock Market Prediction Using Financial News	Schumaker, 2006

In the following section we are going to explain the methodology used by different researchers in various steps of text classification task in stock trend prediction. We also provide some pros and cons related to each article and make overall comparisons among different approaches.

2.3.2 Review of Major Preliminaries

As stated earlier there are many researches related to the impact of public information on stock market variables. But the first systematic examination against the impacts of textual information on the financial markets is conducted by Klein and Prestbo (1974). Their survey consists of a comparison of the movements of Dow Jones Industrial Average with general news during the period from 1966 to 1972. The news stories that they have taken into consideration are the ones appearing in the "What's New" section of Wall Street Journal as well as some featured stories carried on the Journal's front page. The details of news story selection are not mentioned in their work. One of the major criticisms of their study is that too few news stories are taken into account in each day. And stories on the journal's front page are not enough for summarizing and reflecting the information appeared in the whole newspaper. Although with such simple settings they found that the pattern of directional correspondence between the news stories and stock price movements manifested itself 80% of the time. Their findings strongly suggest that news stories and financial markets tend to move together.

The first online system for predicting the opening prices of five stock indices (Dow Jones Industrial Average [Dow], Nikkei 225 [Nky], Financial Times 100 Index [Ftse], Hang Seng Index [His], and Singapore Straits Index [Sti]) was developed by Wuthrich et al. (1998). The prediction is based on the contents of the electronic stories downloaded from the Wall Street Journal. Mostly textual articles appearing in the leading and influential financial newspapers are taken as input. The system is going to predict the daily closing values of major stock markets indices in Asia, Europe, and America. The forecast said to be available real-time via www.cs.ust.hk/~beat/Predict daily at 7:45 a.m. Hong Kong time. Hence predictions would be ready before Tokyo, Hong Kong, and Singapore, the major Asian markets, start trading. News sources containing financial analysis reports and information about world's stock, currency and bond markets are downloaded by the agent. The database named *Today's News*. The latest closing values were also downloaded by the agent and saved in *Index Value*. Old News and Old Index Values contained the training data, the news, and closing values of the last one hundred stock trading days. Keyword tuples contained more than 400 individual sequences of

words provided once by a domain expert and judged to be influential factors potentially moving stock markets. Figure 2.5 presents the prediction methodology used by Wuthrich et al. (1998)

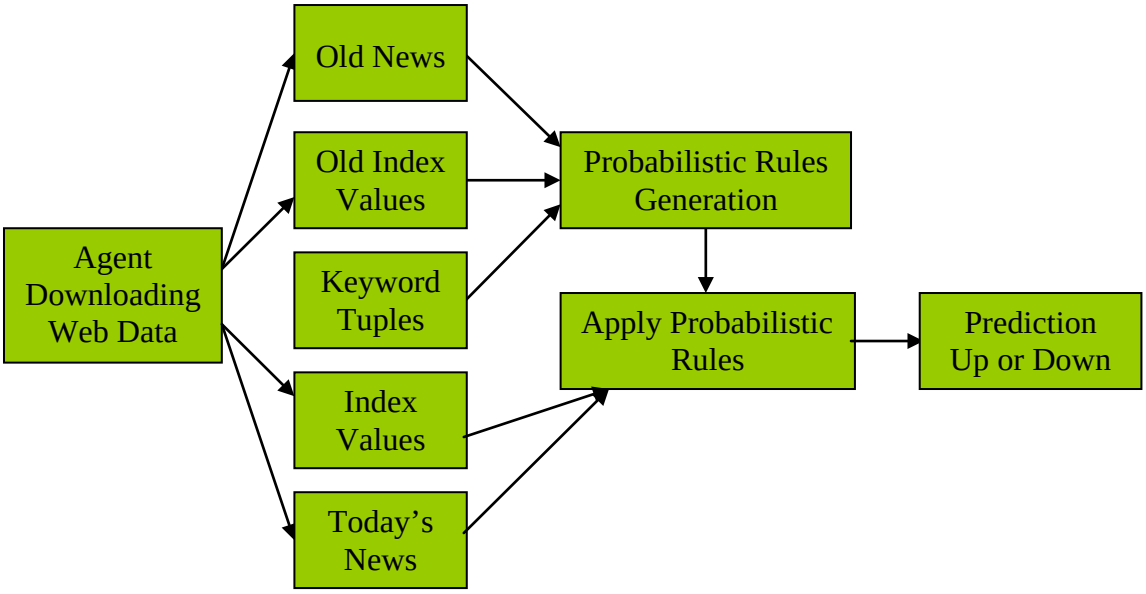


Figure 2.5: Architecture and Main Components of Wuthrich Prediction System
Source: Wuthrich et al., 1998

Various ways of transforming keyword tuple counts into weights have been investigated and several learning techniques, such as rule-based, nearest neighbor and neural net have been employed to produce the forecast. Rule based techniques (Wuthrich, 1995, 1997) proved to be more reliable with higher accuracy than other techniques.

Although the prediction accuracy is significantly above random guessing and their techniques complement numeric forecasting methods as exploiting textual information in addition to numeric time series data increases the quality of result, but there exist some drawbacks allied to this system. First of all the system is just based on keywords provided by domain experts. There might be some new and important words in news articles which are not taken into account and affect the accuracy of results. According to their system, only 5 stock market indices are going to be forecasted and their model is not stock-specific. However stock-specific models have their own problems but financial investors are more interested to have the prediction of each single stock. Thirdly, their input

sources are very limited and should consider others of higher quality. And the last issue is that their system only predicts the opening prices of financial markets and more challenging issues, such as intraday stock price predictions, could not be achieved.

Fawcett and Provost (1999) formulated an *activity monitoring task* for predicting the stock price movements based on the content of the news stories. Activity monitor task is defined as the problem that involves monitoring the behaviors of a large population of entities for interesting events which require actions. The objective of the activity monitoring task is to issue alarms accurately and quickly. In the stock price movement detection, the goal is to scan news stories associated with a large number of companies and to issue alarms on specific companies when their stocks are about to exhibit positive activity. News stories and stock prices for approximately 6,000 companies over three month's period are archived. An interesting event is defined to be a 10% change in stock price which can be triggered by the content of the news stories. The goal is to minimize the number of false alarms and to maximum the number of correctly predicted price spikes. It is worth noting that, the authors only provide a framework for formulating this predicting problem. The implementation details and an in-depth analysis are both missing. Perhaps this is because their main focus is not on examining the possibility of detecting stock price movements based on news stories, but is on outlining a general framework for formulating and evaluating the problems which require continuous monitoring their performance. What can be realized from their work is that they have reduced each news story to a set of constituent stems and stem bi-grams. And one of the limitations of their study is that the alarm is going to be issued only for 10% or greater jumps in stock prices.

Lavrenko et al. (1999, 2000) have done an extensive job in prediction of stock prices based on news articles. They have issued 3 articles with different titles with almost the same content. They have proposed a system called ÁÉnalyst for predicting the intra-day stock price movements by analyzing the contents of the real-time news stories. Analyst is developed based on a language modeling approach proposed by Ponte and Croft (1998). ÁÉnalyst is a system which models the dependencies between news stories and time series. It is a complete system which collects two types of data, processes them,

and attempts to find the relationships between them. The two types of data are financial time series and time-stamped news stories. The system design is presented in Figure 2.6

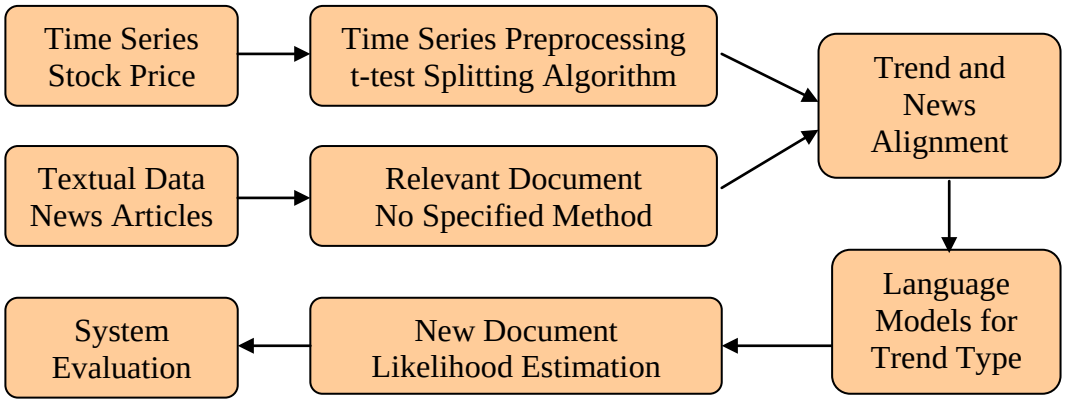


Figure 2.6: Lavrenko System Design

Source: Lavrenko et al. 2000

They have used t-test splitting (Top-Down) algorithm to identify time series trends. They have discretized trends where labels are assigned to segments based on their characteristics including length, slope, intercept, and r^2. This is done using agglomerative clustering algorithm (Everitt, 1993). These labels would be the basis for correlating trends with news stories. Aligning each trend with time-stamped news stories would be the next step and a document is associated with a trend if its time stamp is *h* hour or less before the beginning of trend. They suggest that 5 to 10 hours tends to work best. Language models (Ponte and Croft, 1998; Walls et al., 1999) of stories that are correlated with a given trend are learned. Learning language determines the statistics of word usage pattern among the stories in training set. For evaluating their system, they have used both market simulation and also Detection Error Tradeoff (DET) curves which are similar to ROC curves, common in classification algorithm.

One of the positive aspects of their work is the use of language model which incorporate the entire vocabulary used in the text rather than concentrating on features selected by experts as in work conducted by Wuthrich et al. (1998). Another issue of particular importance is the use of stock specific model. They trained a separate set of models for each stock. The advantage of this model is that it can learn the specific model of language that affects each stock. The main disadvantage of stock-specific models is the

small size of their training set. It means that companies that are rarely covered by news releases are at a disadvantage. Predicting the stock indices solves the problem of shortage in stock news about companies but the models are not able to distinguish the specific effect of news on a particular company.

Lavrenko et al. claim that there should be a period, *t*, to denote the time for the market to absorb any information (news stories) release, where *t* is defined 5 hours in their system. Researchers admit that the market may spend time to digest information, but such a long period may contradict with most economic theories, the efficient market hypothesis. When the system generates profit, it is the sign of market inefficiency. They also argue that with this long time lag, some news stories may classify to trigger both the rise and drop movement of the stock prices in the training stage. Lavrenko himself has admitted such problem and tries to reduce the amounts of overlap by decreasing *h*. In general we can say that their system is capable of producing profit that is significantly higher than random. Many researchers nowadays refer to ÁÉnalyst as a trusted and reliable prediction system.

Thomas and Sycara (2000) predict the stock prices by integrating the textual information that is downloaded from the web bulletin boards into trading rules. The trading rules are derived by genetic algorithms (Allen and Karjalainen, 1995) based on numerical data. For the textual data, a maximum entropy text classification approach (Nigam et al., 1999) is used for classifying the impacts of posted messages on the stock prices. Trading rules are constructed by genetic algorithms based on the trading volumes of the stock concerned, as well as the number of messages and words posted on the web bulletin boards per day. They chose those boards with stocks which were traded on NASDAQ or NYSE as of January first 1999 and those boards with stock prices higher than $1. This left them 22 stocks and the accompanying text from their bulleting boards. They were mostly interested in the profitability of trading rules rather than accuracy of the prediction itself. The authors reported that the profits obtained increased up to 30% by interesting the two approaches rather than using either of them. However no analysis or evaluation on their results is given.

In year 2001, Gidofalvi proposed another model for prediction of stock price movements using news articles. In year 2003 the same article being modified and completed as the technical report with the cooperation of another researcher Charles Elkan (2003). In their articles, they aim to show that short-term stock price movements can be predicted using financial news articles. They also state that a usually less successful technical analysis tries to predict future prices based on past prices, whereas fundamental analysis tries to base predictions on factors in the real economy. Like Lavrenko (2000) and Thomas (2000) their task is rather to generate profitable action signal (buy and sell) than to accurately predict future values of a time series. Figure 2.7 illustrates their prediction system design.

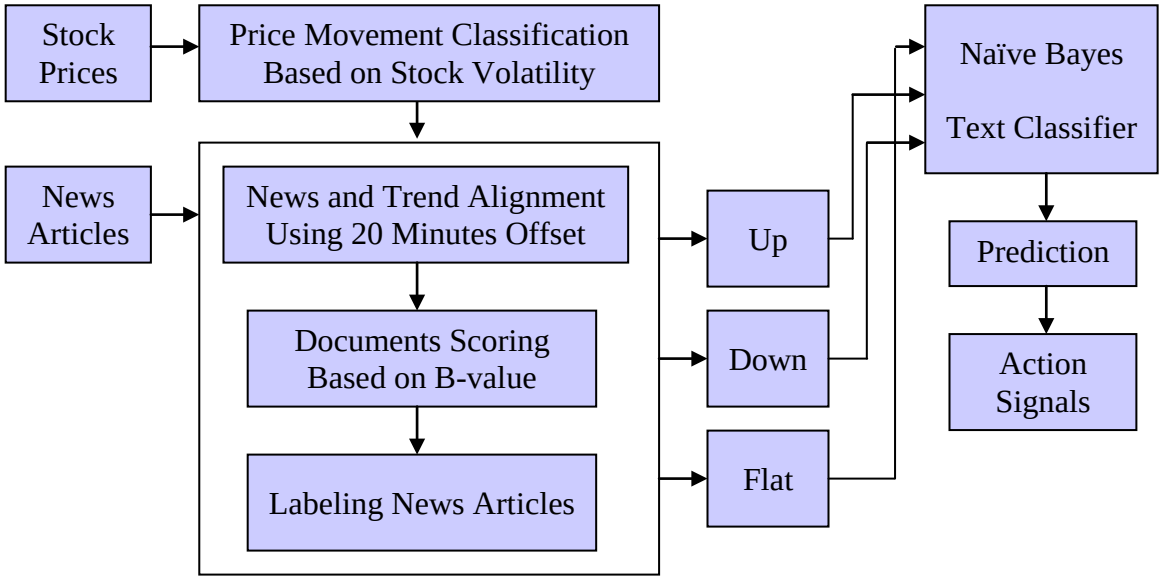


Figure 2.7: Overview of the Gidofalvi System Architecture

Source: Gidofalvi and Elkan, 2003

The first step of their process relates to the identification of movement classes of time series (stock prices) as up, down and unchanged relative to the volatility and the change in a relevant index. For aligning news articles to the movement classes (trends) a time interval is defined which they call the *window of influence*. The window of influence of a news article is the time period throughout which that news article might have an effect on the price of the stock. It is characterized with a lower boundary offset and upper boundary offset from t (news timestamp). They state that in careful experiments,

predictive power of stock price movement is in the interval starting 20 minutes before and ending 20 minutes after news articles become publicly available. It means that they defined the offset 20 minutes. They disregarded news articles that were posted after closing hours, during weekends or on holidays. Scoring of news articles is based on the volatility of stock, which is known as *B*-value. *B*-value describes the behavior or movement of the stock relative to some index, and is calculated using a linear regression on the data points. They scored news articles based on relative movement of the stock price during the window of influence. Stocks with a *B*-value greater than 1 are relatively volatile, while stocks with a *B*-value less than 1 are more stable. Labeling news articles in which each news articles is labeled "up", "down" or "unchanged" is according to the movement of the associated stock in a time interval surrounding publication of the article. Finally they train a Naïve Bayes text classifier for the movement classes. This trained Naïve Bayesian classifier computes the probability for each piece of new stock-specific news articles and identify that particular news article belong to a class representing a particular movement class. They have chosen Rainbow Naïve Bayesian classifier package (McCallum, 1996) for their classification task.

Even though classification results were significant for the [-20, 0] and [0, 20] alignments, the predictive power of the classifier was low. Their result disagrees with the efficient market hypothesis and indicators for the prediction of future stock price behavior can possibly be used in a profitable way. Their model lacks any preprocessing of text including feature selection and extraction. Scoring news articles is based on *B*-value and classification of stock prices is also based on volatility of stocks and changes in relative index. This may not be an incorrect concept, but the more realistic indicators should be taken into account. Because of their particular alignment many news articles have been disregarded which may contain influential matters.

In year 2002, Fung et al., introduces another methodology for stock trend prediction using news articles. A clear distinction of their work is that they want to investigate the immediate impact of news articles on the time series based on the Efficient Market Hypothesis. Building their model based on EMH, a long time lag as in Lavrenko's model in normally impossible. No fixed periods are needed in their system

and predictions are made according to the content of news articles. The overview of their system is shown in Figure 2.8. It consists of two phases: system training phase and operational phase.

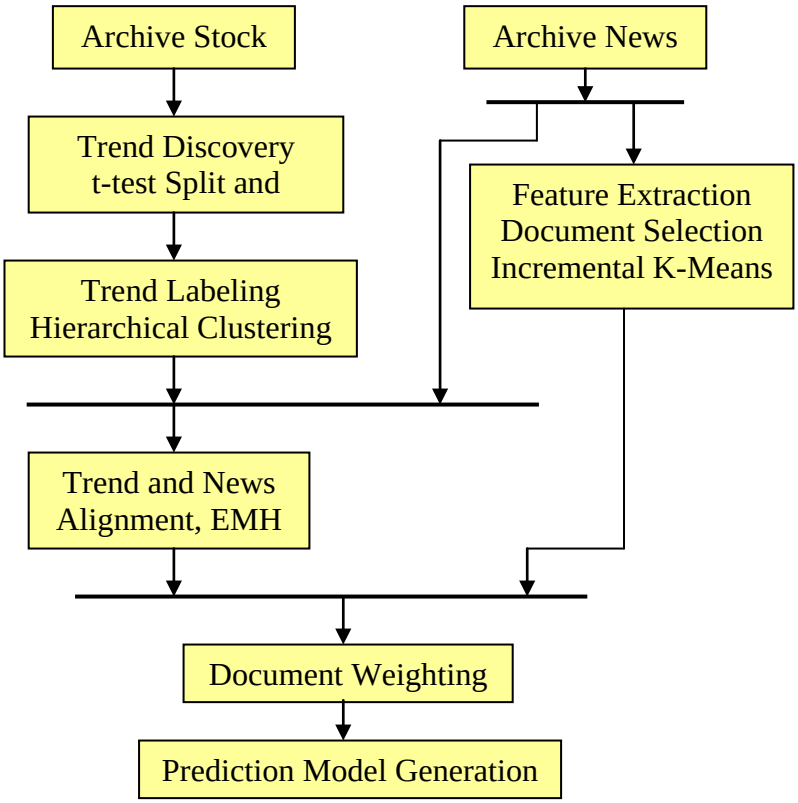


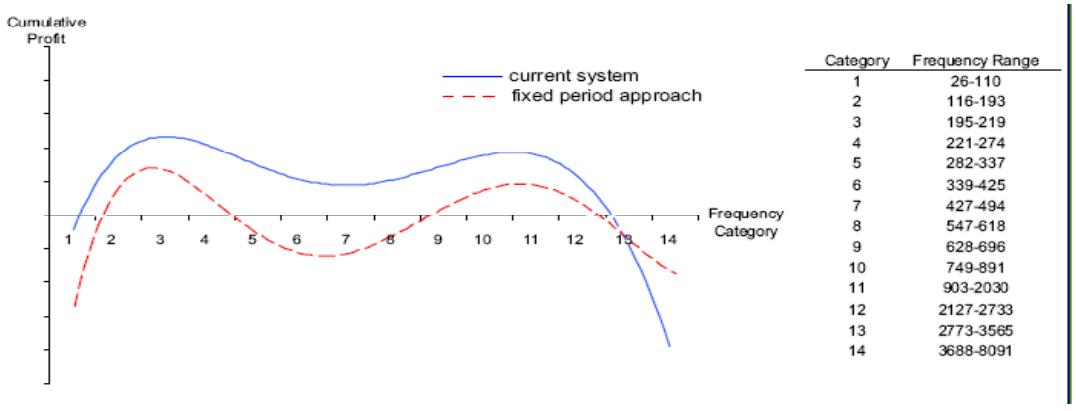
Figure 2.8: An Overview of Fung Prediction Process

Source: Fung et al., 2002

The data and news articles that have been used are 614 stocks in Hong Kong exchange market during 7 consecutive months. Stock trend discovery is based on t-test based splitting and merge algorithm. Using piecewise linear segmentation. Trend labeling which clusters similar segmented trends into two categories, Rise and Drop, is done according to the slope of trends and the coefficient of determination. For this part a two dimensional agglomerative hierarchical clustering algorithm is formulated. Feature selection or useful document selection is handled using a new algorithm named guided clustering, which extracts the main features in the news articles. This algorithm is an extension of the incremental K-Means (Kaufman and Rousseeau, 1990), which can filter

out news articles that do not support the trend. They have chosen K-Means as the recent research findings showed that it outperform the hierarchical approach for textual document clustering (Steinbach et al., 2000, cited by Fung et al., 2002). News articles weighting is based on a differentiated weighting scheme. The association between different features and different trend types are generated based on Support Vector Machine (SVM) (Joachims, 1998). It is a new learning algorithm proposed by Vapnik (1995) and is fully explained in Chapter 4. To evaluate the robustness of the guided clustering algorithm, receiver operating characteristic (ROC) curve is chosen. And for the evaluation of the whole system, a market simulation was conducted and compared to a Buy-and-Hold test.

A comparison between their system and the fixed period alignment approach used by Lavrenko has been made. The underlying assumption between these two models is different. This model is based on EMH, however, the fixed period approach assumes that every related piece of information has an impact on the market after a fixed time interval. The frequency of news articles broadcast must be a critical factor of affecting the prediction performance. Fung et al. (2002) claim that their approach is superior to the fixed period approach as they use all news articles while Lavrenko's approach only uses the news articles within a fixed interval preceding the happening of a trend. And as the frequency of news articles have direct relationship with the profitability of model, their model would be more profitable than fixed period models. (See Figure 2.9)



Category	Frequency Range
1	26-110
2	116-193
3	195-219
4	221-274
5	282-337
6	339-425
7	427-494
8	547-618
9	628-696
10	749-891
11	903-2030
12	2127-2733
13	2773-3565
14	3688-8091

Figure 2.9: Fixed Period vs. Efficient Market Hypothesis; Profit Comparisons

Source: Fung et al., 2002

One year later, in 2003, Fung et al. proposed another predication system mostly the same as the one in 2002, but instead of single time series, they took into account multiple time series. They stated that all the existing approaches are concerned with mining single time series and the interrelationships among different stocks are not well-addressed. They proposed a systematic framework for mining multiple time series based on Efficient Market Hypothesis. The system structure has an additional step, which is the time series relationship discovery and article alignment in multiple time series. Mining multiple time series includes potential related stocks discovery, stock selection using one-tail hypothesis testing on a binomial proportion (Montogometry and Runger, 1999), and article alignment in multiple time series. They have concluded that the profit of monitoring multiple time series is nearly double to that of monitoring single time series. The detailed procedure is not described here as our main focus relies on single time series.

Following the techniques proposed by Wuthrick et al. (1998), Permunetilleke and Wong (2002) repeated the work, but with different a domain. News headlines (instead of news contents) are used to forecast the intra-day currency exchange rate (instead of the opening prices of stock indices). These news headlines belong to world financial markets, political or general economic news. They show that on a publicly available commercial data set, the system produces results are significantly better than random prediction.

In year 2004, Mittermayer proposed a prediction system called NewsCATS (News Categorization and Trading System). It is a system for prediction of stock price trends for the time immediately after the publication of press releases. NewsCATS consists mainly of three components. The first component retrieves relevant information from press releases through the application of text preprocessing techniques. The second component sorts the press releases into predefined categories. Finally, appropriate trading strategies are derived by the third component by means of the earlier categorization. Architecture of NewsCATS is shown in Figure 2.10.

NewsCATS learns a set of categorization rules that allow the Categorization Engine to sort new press releases automatically into a defined number of categories. Each of these categories is associated with a specific impact on the stock prices, e.g., increase

or decrease. Depending on the results yielded by the Categorization Engine (i.e., the category assigned to the new press release) the Trading Engine produces trading signals that can be executed via an online broker or other intermediaries.

Automatic preprocessing of incoming press releases includes feature extraction (parsing, stemming, and stop word removal), feature selection, and document representation. Feature selection is performed by choosing *tf*, *idf*, or *tf*×*idf* as the measure of frequency. Document representation can be performed with a Boolean measure of frequency or with *tf*, *idf*, or *tf*×*idf*.

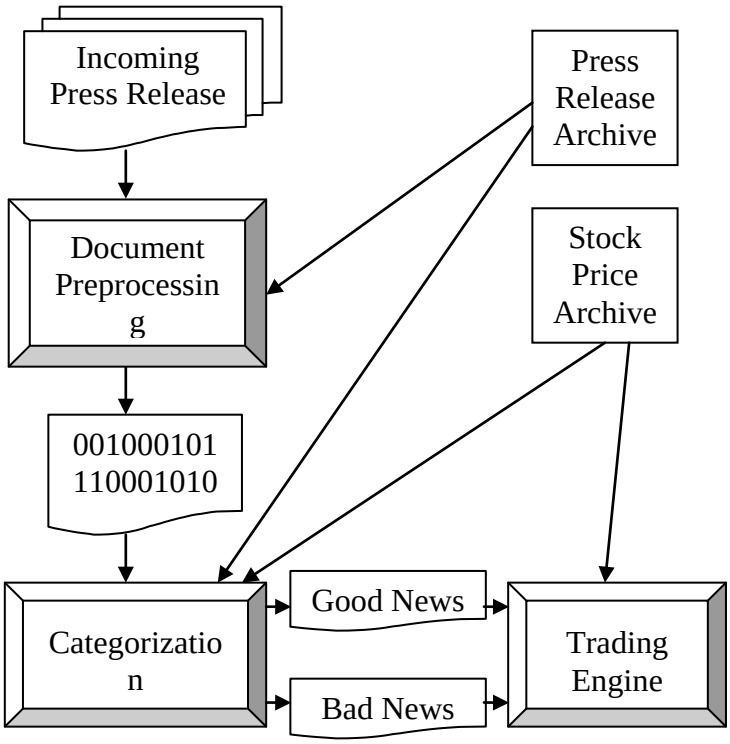


Figure 2.10: Architecture of NewsCATS

Source: Mittermayer, 2004

The output from Preprocessing Engine is forwarded to the Categorization Engine in which the preprocessed documents are categorized into different news types. (Good News or Bad News). The categorization task is implemented using SVM text classifier. On the arrival of a new press release, the host application launches Document

Preprocessing and the Categorization Engine in that order and the Trading Engine generates appropriate trading signals.

Although Mittermayer has proposed an automatic system for the prediction of stock trends, but the details of constructing such a system is missing in his article. Useful document selection and time series preprocessing are not included the model which may reduce the accuracy of the learning system. NewsCATS differs from others mainly in the way that learning examples are chosen and the way the trading recommendations are complied. The output of NewsCATS provides trading strategies that significantly outperform a trader randomly buying and shorting stocks immediately after the publication of press releases. For evaluation of their system, they have used the precision-recall measures and also have conducted a market simulation.

Another research in the area of stock prediction is conducted Khare et al., (2004). They have developed Web news Mining based decisive system for stock movement, named as “Stock Broker P”, i.e. “stock broker prediction”. The classification of the headline is done with the help of Naïve Classifier with modifications. They claim that they have obtained results with the accuracy of over 60%. The main components of their systems includes a technology for extracting small investor sentiment (opinion) and news from the web sources using manipulation of WebPages, opinion mining and sentiment extraction from the web. They create a sentiment index based on available data by aggregating the sentiment value of each news line and they predict the direction of stock market as a whole and of a particular stock by classification based on sentiment index. They have created a dictionary containing significant words by reviewing various stock sites and learning its keywords. Figure 2.11 illustrate “Stock Broker P” structure.

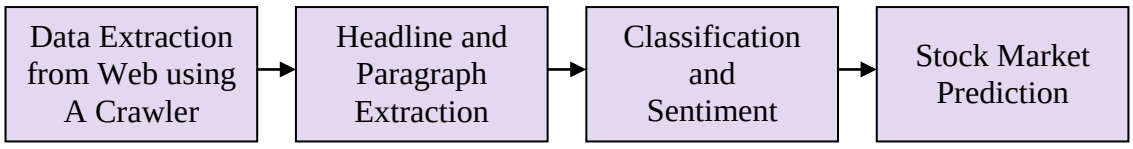


Figure 2.11: “Stock Broker P” System Design

Source: Khare et al., 2004

In year 2005, Fung and Yu with the corporation of another researcher, Lu, proposed another article named “The Prediction Power of Textual Information on Financial Markets”. This is almost the same as the work they proposed in year 2002 with some amendment and modifications. They have recorded the intra-day prices of all the Hong Kong stock and for real-time news stories and more than 350,000 documents are collected. Like previous research, they used a t-test based split and merge segmentation algorithm for figuring out the tertiary movements of stock prices with the difference in the choice of error norm. The alignment process is based on Efficient Market Hypothesis and the selection of useful news unlike the previous work, is based on $\chi 2$ estimation (Refer to Chapter 4) on the keywords distribution over the entire document collection. The selected documents are represented and weighted using the simple *tfidf* model and are normalized to unit length. Finally, the relationship between the contents of the news stories and trends on the stock prices are learned through support vector machines. (Christianini and Shawe-Taylor, 2002) For the evaluation of their work they conducted two market simulations. Simulation 1 is the proposed system and simulation 2 is the Buy-and-Hold Test. Simulation 2 serves as a basis of comparison of their system. They have shown that their system outperforms Simulation 2.

There exists another research in the same domain conducted by Phung (2005). This paper describes a proposed system that extracts key phrases from online news articles for stock movement predictions. The proposed system is implemented and tested on selected active sectors from Bursa Malaysia, a Malaysian Stock Exchange. This paper reviewed investigations on how the online Malaysian news articles are mined to extract appropriate key phrases; which are then subsequently used to predict the movements of stock prices in Bursa Malaysia. The main focus of this paper is to implement appropriate text mining techniques to extract the said key phrases from online news sources. The news obtained from the online sources is processed by HTML Parser that removes HTML tags. They are then fed into the Word Stemmer to strip the suffix and/or prefix of the words in order to reduce the number of words. In feature filter phase the features that provide less information are discarded and this is based on measure frequency such as *tf*, *idf* or *tfidf* method. His proposed system is based on KEA (Witten et al., 1999)

algorithm that automatically extract key phrases from text and employs Naïve Bayes to build a model using a set of training documents.. Precision and recall are used to evaluate the effectiveness of the system. The main focus of his study relies on the key phrase extraction rather than the prediction of stock trend movement.

In 2006, Schumaker and Chen, examine the role of financial news articles on stock trend prediction using 3 different textual representations. They use bag of words, noun phrases and named entities and analyze their abilities to predict discrete stock prices twenty minutes after an article release. Articles and stock quote data are processed by a support vector machine derivative, Sequential Minimal Optimization style of regression which can handle discrete number analysis. They have shown that their model has a statistically significant impact on predicting future stock prices compared to linear regression. They have also demonstrated that using a Named Entities representation scheme performs better than the standard bag of words. The main difference of their proposed system with the other systems relies on the use of noun phrases and named entities in textual representation. They also try to forecast the discrete stock price instead of predicting the trend of the stock movement. They have not mentioned the details of their feature selection and document representation.

2.4 Chapter Summary

Different models for prediction of stock market based on content of news articles have been described in the preliminary works. The number of proposed models is not comparable with the number of models generated for the prediction of stock prices based on structured data. But, researchers have been successful to some extent, in prediction of stock prices based on unstructured data. Most of them have proven that the trading rules generated from their models are more profitable than a random walk trade. They are using various techniques and approaches to guarantee higher accuracy for their prediction. Actually there exist no fixed model for this purpose and many different attributes are involved which affect the accuracy of prediction including the type of data, the assumption used, the text preprocessing, and the type of classifier used. It seems that the future of such prediction models would be bright.

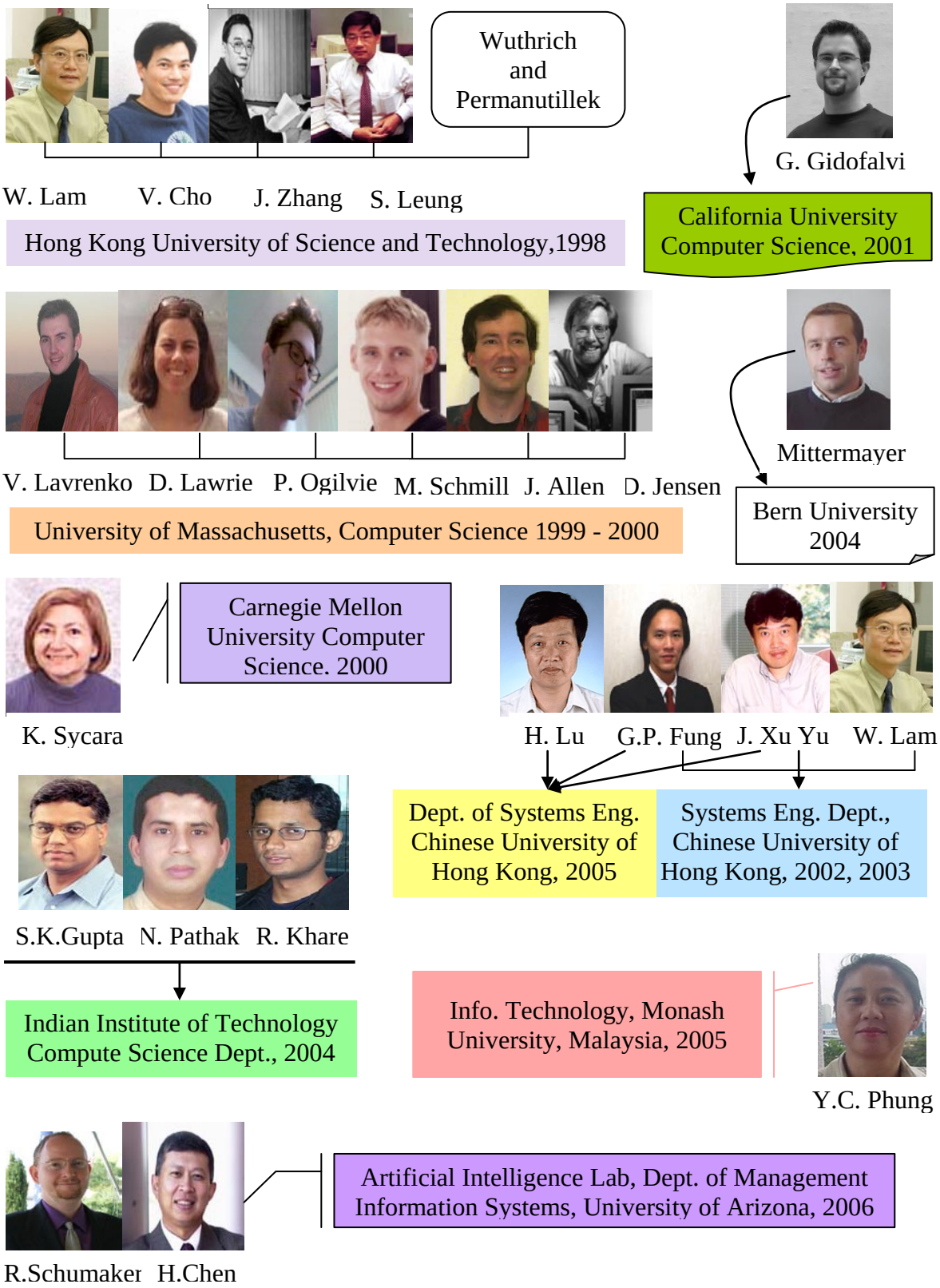


Figure 2.12: Knowledge Map; Scholars of Stock Prediction Using News Articles

Chapter 3

Time Series Preprocessing

3. Time Series Preprocessing

Time-series data is frequently encountered in real world applications. Stock price and economic growth are the typical examples. A time-series represents a set of consecutive observations or measurements taken at certain time intervals. The observations can be real numerical values, for instance, a stock price, or categories, for instance, medicines taken by a patient over a treatment period. These continuous values of a time-series can be converted to categories if needed. Since all stock time series contains a high level of noise, high level time series segmentation is necessary for recognizing the significant movements. Discovery of interesting patterns from a large number of time-series is required in many applications including economic and science.

3.1 Time Series Data Mining

In recent years, there has been a lot of interest within the research community in the mining of time series data, which arise in business as well as scientific decision-support applications; examples include stock prices or currency exchange rates collected over time. (Guo et al., 2002) Finding periodic patterns in time series databases is an important data-mining task with many applications. In time series data mining, periodic pattern help to find the rise and drop of stock values. Loether and McTavish (1993) have

developed many methods for searching periodicity patterns in large data sets. However, most previous methods on periodicity search are on mining *full periodic patterns*, where every point in time contributes (precisely or approximately) to the cyclic behavior of the time series. A useful related type of periodic patterns, called *partial periodic patterns*, which specify the behavior of the time series at some but not all points in time, have not received enough attention. Thus, *partial periodicity* is a looser kind of periodicity than *full periodicity*, and it exists ubiquitously in the real world. (Han et al., 1999)

3.1.1 On Need of Time Series Data Mining

Recently, the increasing use of temporal data has initiated various research and development efforts in the field of data mining. Chung et al. (2002) consider time series as an important class of temporal data objects, which can be easily obtained from financial and scientific applications. They claim that they are major sources of temporal databases and, undoubtedly, discovering useful time series patterns is of fundamental importance. They compare transactional databases with discrete items and time series data, which are characterized by their numerical and continuous nature. Multiple time series (stock data) are difficult to process, analyze and mine. However, when they can be transformed into meaningful symbols like technical patterns, it becomes an easier task.

The sheer volume of data collected means that only a small fraction of the data can ever be viewed hence application of classic machine learning and clustering algorithms on time series data have not met with great success. This is due to the typically high dimensionality of time series data, combined with the difficulty of defining a similarity measure appropriate for the domain. (Keogh and Pazzani, 1998)

It is suggested by Shatkay and Zedonic (1996) to divide the sequences into meaningful subsequences and represent them, using real-valued functions. There is a need to discretize a continuous time series into meaningful labels/symbols. (Das et al., 1998; Kai et al., 1999) This process is called “numeric-to-symbolic” conversion, and should be considered as one of the most fundamental components in time series data mining systems. (Chung et al., 2002) Since the data sets occurring in practice tend to be

very large, most of the works have focused on the design of efficient algorithms for various mining problems. (Guo et al., 2002)

3.1.2 Major Tasks in Time Series Data Mining

One of the tasks for time-series data mining is search for periodic patterns. Yu et al. (2001) address three different kinds of pattern namely cyclical pattern, trend pattern and seasonal pattern. A cyclical pattern exists when the data exhibit rises and falls that are not of a fixed period. A trend pattern exists when there is a long-term increase or decrease in the data. And a seasonal pattern exists when a series is influenced by seasonal factors. Many time-series in the real world have trends and the stock value of an equity fund could be an example of trend pattern. Keogh and Kasetty (2002) have defined the major tasks in time series data mining as followings:

Indexing: Given a query time series Q, and some similarity/dissimilarity measure D(Q,C), find the nearest matching time series in database DB.

Clustering: Find natural groupings of the time series in database DB under some similarity/dissimilarity measure D (Q, C).

Classification: Given an unlabeled time series Q, assign it to one of two or more predefined classes.

Segmentation: Given a time series Q containing n data points, construct a model from K piecewise segments (K << n) such that Q closely approximates Q.

3.2 Time Series Representation

As with most computer science problems, representation of the data is the key to efficient and effective solutions. (Keogh et al., 2001a) According to Keogh and Pazzani (1998), a representation should allow efficient computation on the data, which extracts higher order features. Several such representations have been proposed, including Fourier Transformations (Faloutsos et al., 1994), relational trees (Shaw and deFigueiredo, 1990)

and envelope matching/R+ trees (Agrawal et al., 1995). The above approaches have all met with some success, but all have shortcomings, including sensitivity to noise, lack of intuitiveness, and the need to fine-tune many parameters. (Keogh and Pazzani, 1998)

3.2.1 Piecewise Linear Representation (PLR)

According to Keogh et al. (2001a) one of the most commonly used representations is piecewise linear approximation. This representation has been used by various researchers to support clustering, classification, indexing, and association rule mining of time series data. Piecewise Linear Representation (PLR) also called piecewise linear approximation or segmentation, refers to the approximation of a time series T, of length n, with K straight lines. Because K is typically much smaller than n, this representation makes the strong transmission and computation of the data more efficient. (Keogh et al., 2001a) Piecewise linear representation of time series data is a good technique to reduce the complexity of the raw data. (Ge, 1998) According to Wang C., and Wang X.S. (2000) approximation usually saves disk storage space and hence reduces disk access time, which often is the bottleneck of the searching process.

Piecewise linear segmentation, which attempts to model the data as sequences of straight lines (Figure 3.1) has innumerable advantages as a representation. (Keogh and Pazzani, 1998) Pavlidis and Horowitz (1974) describe piecewise approximation as a way of feature extraction, data compaction, and noise filtering. Keogh (1997) declares that this representation has many desirable properties, including:

- High rates of data compression (Converting many data points to few segments)
- Relative insensitivity to noise
- Intuitiveness and ease of visualization

Data compression is of particular importance. It captures most of the essential information, although it compresses the raw data. It results in computational speedup and allows us to search huge time series in main memory without having to swap in and out small subsections.

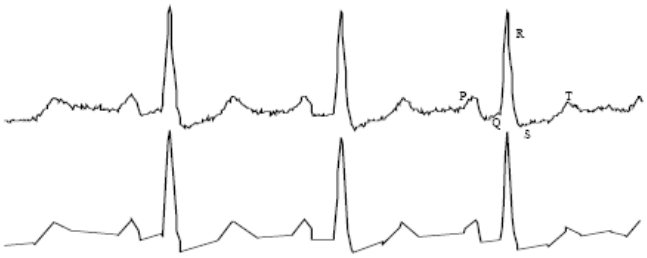


Figure 3.1: Examples of a Time Series and its Piecewise Linear Representation
Source: Keogh and Pazzani, 1998

3.2.2 PLR applications in Data Mining Context

As stated in different papers and literatures, piecewise linear representation has been used extensively in different fields of data mining. Keogh et al. (2001a) have summarized the main application of piecewise linear representation in the data mining context. They state that PLR is able to:

- Support fast exact similarly search (Keogh et al., 2001b)
- Support concurrent mining of text and time series (Lavrenko et al., 2000)
- Support novel clustering and classification algorithms (Keogh et al, 1998)
- Support change point detection (Ogden and Sugiura, 1994)

As stated by Guo et al. (2002) a common task with a time series database is to look for an occurrence of a particular pattern within a longer sequence. According to Wang C. and Wang X.S. (2000), in technical analysis of time series, especially stock market analysis, the study of shapes is very important. The time series is drawn in a two-dimensional plane, where one dimension is the time and the other is the value. The shape of a time series is usually taken as the shape of the curve formed by connecting consecutive points on the plane.

Piecewise linear approximation of such plane can be used to identify patterns associated with growth in stock prices or to identify non-obvious relationships between two time series. (Guo et al., 2002)

3.2.3 Piecewise Linear Segmentation algorithms

Time series are ubiquitous in scientific research, and accurate segmentation of time series is essential to the performance of many applications. Time series segmentation is important in many of these areas because it is often used as a preprocessing step in time series analysis applications. (Bouchard, n.d.) There are numerous algorithms available for segmenting time series, many of which where pioneered by Pavlidis and Horowitz (1974). Keogh et al. (2001a) refer to any types of algorithm, which input a time series and return a piecewise linear representation, as segmentation algorithms. Although appearing under different names and with slightly different implementation details, most time series segmentation algorithms can be grouped into one of the following three categories. Keogh et al. (2001a) give a thorough overview of these techniques:

Sliding Windows: a segment is grown until it exceeds some error bound. The process repeats with the next data point not included in the newly approximated segment. The Sliding Windows algorithm works by anchoring the left point of a potential segment and the first data point of a time series, then attempting to approximate the data to the right with increasing longer segments. At some point *i*, the error for the potential segment is greater than the user-specified threshold, so the subsequences from the anchor to $i - 1$ is transformed into a segment. The anchor is moved to location *i*, and the process repeats until the entire time series has been transformed into a piecewise linear approximation. The Sliding Windows algorithm is attractive because of its great simplicity and intuitiveness.

Top-Down (Split): The time series is recursively portioned until some stopping criterion is met. The Top-Down algorithm works by considering every possible partitioning of the time series and splitting it at the best location. Both subsections are then tested to see if their approximation error is below some user-specified threshold. If not, the algorithm recursively continues to split the subsequences until all the segments have approximation errors below the threshold. Lavrenko et al. (2000) use the Top-Down algorithm to support the concurrent mining of text and time series to investigate the

influence of news stories on financial markets. Their algorithm contains some interesting modification including a novel stopping criterion based on the t-test.

Bottom-Up (Merge): Starting from the finest possible approximation, segments are merged until some stopping criteria are met. The Bottom-Up algorithm is the natural complement to the Top-Down algorithm. The algorithm begins by creating the finest possible approximation of the time series, so that n/2 segments are used to approximate the n length time series. Next, the cost of merging each pair of adjacent segments is calculated, and the algorithm begins to interactively merge the lowest cost pair until a stopping criterion are met. When the pair of adjacent segments i and $i+1$ are merged, the algorithm needs to perform some bookkeeping. First, the cost of merging the new segment with its right neighbor must be calculated. In addition, the cost of merging the $i - 1$ segment with its new larger neighbor must be recalculated. In data mining, the algorithm has been used extensively by Keogh and Pazzani (1998, 1999) to support a variety of time series data mining.

Durell Bouchard (n.d.) gives a comparative study on the different time series segmentation algorithm. According to his study, the Sliding Windows is extremely slow and not useful for real-time applications. If finding the best segmentation is not imperative to an application, it can be made faster by limiting the size of the sliding window and by increasing the amount the window slides. Top-Down performance is dependent on the number of segments produced, and therefore, the distance measure used. This method is faster than the Sliding Window method, but it is still slow. It is, therefore, not appropriate for real-time applications. Like the Top-Down similarity method, performance of Bottom-Up is dependent on the distance measure. However, this segmentation method is used more frequently than the Top-Down method because it produces similar results and is faster.

3.2.4 Linear Interpolation vs. Linear Regression

When approximating a time series with straight lines, there are at least two ways we can find the approximating line: (Keogh et al., 2001a)

Linear Interpolation in which the approximating line for the subsequence T [a:b] is simply the line connecting t_a and t_b. This can be obtained in constant time.

Linear Regression in which the approximation line for the subsequence T [a:b] is taken to be the best fitting line in the least square sense. (Shatkay, 1995) This can be obtained in time linear in the length of segment.

Linear interpolation tends to closely align the endpoint of consecutive segments, giving the piecewise approximation a "smooth" look. In contrast, piecewise linear regression can produce a very disjointed look on some datasets. (Figure 3.2) The aesthetic superiority of linear interpolation, together with its low computational complexity has made it the technique of choice in many domains.

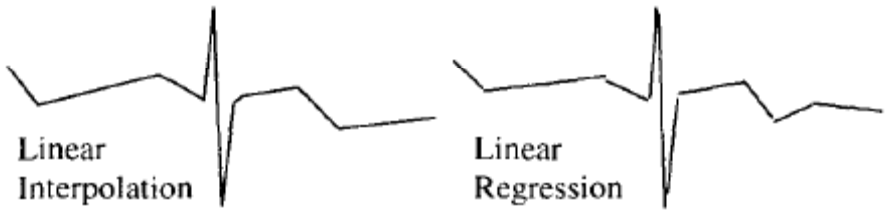


Figure 3.2: Linear Interpolation vs. Linear Regression
Source: Keogh et al., 2001a

3.2.5 Stopping Criterion and the Choice of Error Norm

One of the important issues in using segmentation algorithms relates to the number of segments to be produced. (Sripada et al., 2002) All segmentation algorithms need a stopping criterion to terminate iteration. In its most common form given by Keogh et al (2001a), segmentation problem is framed in several ways:

1. Given a time series T, produce the best representation using only K segments.
2. Given a time series T, produce the best representation such that the maximum error for any segment does not exceed some user-specified threshold, max_error.
3. Given a time series T, produce the best representation such that the combined error of all segments is less than some user-specified threshold, total_max_error.

Keogh (1997) points out the importance of choosing the right number of linear segments used in approximating a given time series. If we choose too small a K, we will lose important features and may fail in our goal of similarity search. On the other hand, if we choose too large a K, we capture all the necessary detail in our time series, but retain redundant information. This is of particular importance because the speedup achieved by the search algorithm presented later in this paper is proportional to the *square* of the compression rate.

The performance of the algorithms depends on the value of max_error. (Keogh et al., 2001a) As max_error, goes to zero all the algorithms have the same performance, since they would produce n/2 segments with no error. At the opposite end, as max_error becomes very large, the algorithms once again will have the same performance, sine they all simply approximate T with a single best-fit line. Instead, a relative performance for some reasonable value of max_error should be tested, a value that achieves a good trade off between compression and fidelity. This “reasonable value” is subjective and dependent on the data mining application and the data itself.

All segmentation algorithms need some method to evaluate the quality of fit for a potential segment. As stated by Keogh et al. (2001a) a measure commonly used in conjunction with linear regression is the sum of squares, or the residual error. This is calculated by taking all the vertical differences between the best-fit line and the actual data points, squaring them, and then summing them together. Another commonly used measure of goodness of fit is the distance between the best-fit line and the data point furthest away in the vertical direction.

Pavlidis and Horowitz (1974) two of the pioneers in time series segmentation have given a literature on the choice of error norm in their paper "Segmentation of Plane Curve". They have stated that the choice of the error norm depends to some extent on the type of data and its application. The error norm over a point set S_k can be defined in many ways. For example e_i can be defined as the Euclidean distance between $\{x_i,y_i\}$ and the approximating curve evaluated at that point so the error norm with respect to the regression line would be $E_i = | \sin\theta.x_i + \cos\theta.y_i - d |$. Two of the most common error norms are integral square error E_2, where $E_2 = \Sigma e_i^2$ for (xi,yi) ϵ S_k and maximum error $E\infty = \max (e_i)$ for (xi,yi) ϵ S_k. Maximum error results in approximation which present a closer fit but involve more computational effort.

3.2.6 "Split and Merge" Algorithm

Split and merge segmentation algorithm includes both the Top-Down and Bottom-Up techniques. This algorithm has been used extensively in different applications such as graphics and image processing. (Wu, 1993; Yang and Lee, 1997; Borges and Aldon, 2000) Pavlidis and Horowitz (1974) have explained the application of split and merge algorithm in segmentation of plane curves. They state that the breakpoint adjustment can be accelerated substantially if one is allowed to merge adjacent segments with similar approximating coefficients and split segments with large error norms. In this way, it is easy to obtain segmentations where the error norm on each segment (or over all of them) does not exceed a specified bound. Merging of segments as described above results in controlled increases of the error norm. They have defined this segmentation algorithm as a 2-step procedure. S_l, S_2,….,S_{nr} is defined to be the segments at the r^{th} iteration and E_l, E2,…,E_{nr} the respective error norms.

Step 1: For i = 1, 2,…,n, check if E_i exceeds E_{max}. If it does, split the i^{th} interval into two and increment n_r. The dividing point can be determined by different rules. Then calculate the error norms on each new interval. **Step 2:** For i = 1, 2... , n - 1 merge segments S_i and S_{i+1}, provided that this will result in a new segment with $E_i < E_{max}$. Then decrease n, by one and calculate the error norm on the new interval.

Pure merging algorithms are computationally expensive. When using split and merge algorithm there is no need to start merging from the small initial individual points as the splitting phase have reduced the number of data points. The split-and-merge algorithm removes the need for a careful choice of the number of segments at the first place. In addition, it results in faster curve fitting than other schemes. (Morse, 2000)

The split and merge algorithm have been used in time series segmentation problem by different researchers. Fung et al. (2002, 2003, and 2005) have used split and merge algorithm for figuring out the tertiary movements of stock prices. Lavrenko et al. (1999, 2000) use the segmentation algorithm to support the concurrent mining of text and time series to investigate the influence of news stories on financial markets.

3.3 Summary

Time series are found in many different areas of scientific study including chemistry, physics, geology, and economics. Time series segmentation is important in many of these areas because it is often used as a preprocessing step in time series analysis applications. When a time series is segmented manually, it is not only extremely slow but it is also inconsistent. Inconsistent segmentation can have adverse affects on the performance of some time series analysis systems. (Bouchard, n.d.)

Most recent work on time series queries only concentrates on how to identify a given pattern from a time series. Researchers do not consider the problem of identifying a suitable set of time points for segmenting the time series in accordance with a given set of pattern templates (e.g., a set of technical patterns for stock analysis.) Using fixed length segmentation is a primitive approach to this problem; hence, a dynamic approach is preferred so that the time series can be segmented flexibly and effectively according to the needs of the users and the applications. (Chung et al., 2002)

By reviewing the literature, one can realize the progress, which has been made in time series linear approximation. Researchers have proposed different segmentation algorithms with different scientific applications. These algorithms play an important role in time series preprocessing and make time series analysis practical and more efficient.

Chapter 4

Literature on Text Categorization Task

4. Literature on Text Categorization Task

Text categorization is a machine learning task with the problem of automatically assigning predefined categories to free text documents. While more and more textual information is available online, effective retrieval is difficult without good indexing and summarization of document contents. TC is now being applied in many contexts, ranging from document indexing based on a controlled vocabulary, to document filtering, automated metadata generation, word sense disambiguation, population of automatic hierarchical catalogues of Web resources, and in general any application requiring document organization or selective and adaptive document dispatching.

4.1 Synopsis of Text Categorization Problem

The automated categorization (or classification) of texts into predefined categories has witnessed a booming interest in the last 10 years, due to the increased availability of documents in digital form and the ensuing need to organize them. In the research community the dominant approach to this problem is based on machine learning techniques: a general inductive process automatically builds a classifier by learning, from a set of pre-classified documents, the characteristics of the categories. The advantages of this approach over the knowledge engineering approach (consisting in the manual

definition of a classifier by domain experts) are a very good effectiveness, considerable savings in terms of expert labor power, and straightforward portability to different domains. (Sebastiani, 2002)

Text Categorization dates back to the early ’60s, but only in the early ’90s did it become a major subfield of the information systems discipline. It is the automated assigning of natural language texts to predefined categories based on their content. (Lewis, 1992) Text categorization is a machine learning task and is of increasing importance. (Zheng and Srihari, 2003)

Formally, TC consists of determining whether a document d_i (from a set of documents *D*) belongs or not to a category c_j (from a set of categories *C*), consistently with the knowledge of the correct categories for a set of train documents. (Sebastiani, 2002)

4.1.1 Importance of Automated Text Categorization

As the volume of information available on the Internet and corporate intranets continues to increase, there is a growing need for tools helping people better find, filter, and manage these resources. Text classification is an important component in many information management tasks. However, with the explosive growth of the web data set, algorithms that can improve the classification efficiency while maintaining accuracy, are highly desired. (Wang et al., 2006)

Automatic text categorization becomes more and more important for dealing with massive data such as digital libraries, news sources, and inner data of companies which are surging more and more. With the advent of WWW (World Wide Web), text categorization becomes a key technology to deal with and organize large numbers of documents. More and more methods based on statistical theory and machine learning has been applied to text categorization in recent years. (Shang et al., 2006)

The categorization task has typically been done by human experts. However, as the number of texts increases, it becomes difficult for humans to consistently categorize

them. Human categorization is very time-consuming and costly and thus its applicability is limited especially for very large document collections. The inability of users to assimilate the profitability utilize such large numbers of documents becomes more and more apparent. Therefore, automatic text categorization is a successful paradigm for organizing documents and an essential technology for intelligent information systems, which has received much attention in recent years. (Tokunaga and Iwayama, 1994; Masuyama and Nakagawa, 2002) Text classification techniques have increased in importance and economic value for digital world as they develop key technologies for classifying new electronic documents, finding interesting information on web and guiding a user's search through hypertext. (Fragos et al., 2005)

4.1.2 Text Categorization Applications

Text categorization is an important research area in many Information Retrieval (IR) applications. Many information retrieval problems such as filtering, routing, or searching for relevant information, benefit from the text categorization research. (Yan et al., 2005) It has wide applications such as email filtering, information organization, document routing and etc. (Doan and Horiguchi, 2004a) Filtering is defined as the absolute assessment meaning that if "Document d" is relevant and routing is defined as the relative assessment meaning that if "Document d1" is more relevant than "d2". (Manomaisupat and Abmad, 2005) Machine learning for text classification is the cornerstone of document categorization, news filtering, document routing, and personalization. The potential is great for machine learning to categorize, route, filter, and search for relevant information. (Forman, 2003) According to Lewis (1992) text categorization applications include indexing texts to support document retrieval and extracting data from texts.

Automatic text categorization can play an important role in a wide variety of more flexible, dynamic and personalized information management tasks such as: real-time sorting of email or files into folder hierarchies; topic identification; structured search and/or browsing; finding documents that match long-term standing interests or more dynamic task-based interests.(Dumais et al. 1998)

4.1.3 Text Categorization General Process

Text categorization task involves many different steps and processes. Researchers have classified these processes in a different way and under different names. In this section, we have reviewed some of the ways researchers have defined the TC process.

Montanes et al. (2003) define text categorization process as document representation, feature reduction, and classification. Debole and Sebastiani (2003) point out that the construction of a text classifier involves 3 phases: A phase of *term selection*, in which the most relevant terms for the classification task are identified. A phase of *term weighting*, in which document weights for the selected terms are computed and a phase of *classifier learning*, in which a classifier is generated from the weighted representations of the training documents. Term selection and term weighting may come under a single phase called *document indexing* which means the creation of internal representations for documents.

Apte et al. (1994) states the 4 different phases that should be implemented for text categorization: **Preprocessing step:** for determining the values of the features or that will be used for representing the individual documents within a collection. **Representation step:** for mapping each individual document into training sample using the above dictionary, and associating it with a label that identifies its category. **Induction step:** for finding patterns that distinguish categories from one another. **Evaluation step:** for choosing the best solution, based on minimizing the classification error or cost.

The state-of-the-art classification framework should be viewed as the process of text representation, classifier training, and performance evaluation. (Wang Y. and Wang X.J., 2005) He classifies text representation process into three phases including text preprocessing, feature reduction, and document representation. In part 4.2 we review the first step in text categorization process namely text preprocessing. Feature reduction will be discussed in part 4.3. Document representation and classifier learning would be explained in part 4.4 and 4.5 accordingly.

4.2 Text Preprocessing

As given by Wang Y. and Wang X.J. (2005) pre-processing is to make clear the border of each language structure and to eliminate as much as possible the language dependent factors. It consists of different tasks, the importances of which are tokenization, stop-word removal, and stemming. Lee and Chen (2006) define the task of preprocessing as removing stop words and word stemming.

In computer science, tokenization is the process of demarcating and possibly classifying sections of a string of input characters. The resulting tokens are then passed on to some other form of processing. (Wikipedia, the Free Encyclopedia, 2001a)

According to Wen et al. (2003) token identification task is to automatically identify the boundaries of a variety of phrases of interest in raw text and mark them up with associated labels. The initial task is to produce a list of attributes from samples of text of labeled documents, i.e., the dictionary. These attributes could be single words or word phrases such as bi-grams or n-grams. (Apte et al., 1994)

Bi-grams are groups of two written letters, two syllables, or two words, and are very commonly used as the basis for simple statistical analysis of text, and are one of the most successful language models for speech recognition. They are a special case of n-gram. An n-gram is a sub-sequence of *n* items from a given sequence. n-grams are used in various areas of statistical natural language processing. An *n*-gram of size 2 is called "bi-gram". (Wikipedia, the Free Encyclopedia, 2001b)

The number of features can be dramatically reduced by the domain dependent methods which include the elimination of stop words, stripping of special characters as well as stemming algorithms or morphological analysis. (Novovicova and Malik, 2005) According to Liu et al. (2005) the feature dimensionality can be cut down by removing stop-words and words with high frequency. Stop words are usually given as a word list. Most of these words are conjunctions or adverbs which have no contribution to classification process, and sometimes have negative influence. Words with high frequency which appear in most documents are not helpful for classification either.

Words appear in no more than three documents and at least 33% of all documents can be removed. Forman (2002, 2003) states that common words can be identified either by a threshold on the number of documents the word occurs in, e.g. if it occurs in over half of all documents, or by supplying a *stop word* list. Stop words are language-specific and often domain-specific.

It is also to be mentioned that the common practice of *stemming* or *lemmatizing* also reduces the number of features to be considered. It is the process of merging various word forms such as plurals and verb conjugations into one distinct term. (Forman, 2002) Song et al., (2005) state that the words in the dictionary that share the same morphological root may be merged. Supporters for word stemming argue that it can not only reduce the dimensionality of feature space but also be helpful to promote the effectiveness of a text classifier.

Various algorithms have been proposed by different researchers for implementation of stop word removal and stemming. As these algorithms are not the purpose of this study, their literatures are not brought at this point.

4.3 Dimension & Feature Reduction Techniques

Manomaisupat and Abmad (2005) define dimensionality reduction as the exclusion of a large number of keywords, based preferably on a statistical criterion; to create a low-dimensional vector. Dimension Reduction techniques have attracted much attention recently since effective dimension reduction make the learning task such as categorization more efficient and save more storage space. (Yan et al., 2005)

The documents in text categorization are represented by a great amount of features and most of them could be irrelevant or noisy. (Montanes et al., 2003)

The set of features is still too large for many learning algorithms such as neural networks even after the elimination of stop words and performance of stemming (domain dependent methods). In order to improve scalability of text categorization, feature reduction techniques should be applied to reduce the feature size further more using

domain independent methods. (Zheng and Srihari, 2003) Dimension reduction techniques can generally be classified into Feature Extraction (FE) approaches (Liu and Motoda, 1998) and Feature Selection (FS) approaches which will be discussed in the following section.

4.3.1 Feature Selection vs. Feature Extraction

It is very important to make a distinction between different approaches of dimensionality reduction. Many researchers have defined and studied these methods which will be reviewed in this section. For the negative influence of high dimensionality and data sparsity, it is highly desirable to reduce the feature space dimensionality. There are two commonly used techniques to deal with this problem: feature extraction (re-parameterization) and feature selection. (Schütze et al., 1995)

Novovicova and Malik (2005) classify these approaches as feature selection and feature extraction because of the nature of resulting features: dimensionality reduction by feature selection in which the feature set is a subset of the original feature set or dimensionality reduction by feature extraction in which the reduced feature set is not of the same type of the features in original feature set but are obtained by combinations or transformations of the original feature variables.

Feature Selection (FS) is one of the feature reduction approaches most used in text categorization. The main idea is to select a subset of features from the original one. FS is performed by keeping the words with highest score according to a predetermined measure of the importance of the word. (Montanes et al., 2003) The selected feature retains original physical meaning and provides a better understanding for the data and learning process. (Liu et al., 2005)

Wyse et al. (1980) define feature extraction as a process that extracts a set of new features from the original features through some functional mapping such as principal component analysis (PCA) (Jolliffe, 1986; Smith, 2002), Latent Semantic Indexing (Deerwester et al., 1990; Papadimitriou et al., 1998) and random projection (Kaski, 1998; Bingham and Mannila, 2001; Fradkin and Madigan, 2003; Deegalla and Bostrom, 2006;

Lin and Gunopulos, 2003; Achlioptas, 2001; Vempala, 1998; Fern and Brodley, 2003) There are also some comparative studies on the dimension reduction techniques. (Fodor, 2002; Tang et al., 2004; Vinay et al., 2005) This method has a drawback that the generated new features may not have a clear physical meaning so that the clustering results are difficult to interpret. (Dash and Liu, 2000; Liu et al., 2005) Though the FE algorithms have been proved to be very effective for dimension reduction, the high dimension of data sets in the text domain often fails many FE algorithms due to their high computational cost. Thus FS algorithms are more popular for real life text data dimension reduction problems. (Yan et al., 2005)

4.3.2 Importance of Feature Selection in Text Categorization

A major characteristic, or difficulty, of text categorization problems is the high dimensionality of the feature space. The native feature space consists of the unique terms (words or phrases) that occur in documents, which can be hundreds or thousands of terms for even a moderate-sized text collection. This is prohibitively high for many learning algorithms. Few neural networks, for example, can handle such a large number of input nodes. Moreover most of these dimensions are not relative to text categorization; even some noise data hurt the precision of the classifier. Hence it is highly desirable to reduce the native space by selecting some representative features from the original feature space in order to improve the efficiency and precision of classifier without sacrificing categorization accuracy. It is also desirable to achieve such a goal automatically, i.e., no manual definition or construction of features is required. (Yang and Pedersen, 1997; Shang et al., 2006; Wang et al., 2006)

One of the most interesting issues in machine learning in general and text categorization in particular is feature selection which selects “good” features for a classifier. (Doan and Horiguchi, 2004a) The aim of feature selection methods is to reduce the dimensionality of a vector. (Manomaisupat and Abmad, 2005) The task of automatic feature selection methods includes the removal of non-informative terms according to corpus statistics and the construction of new features. (Yang and Pedersen, 1997; Soucy and Mineau, 2003)

According to Forman (2002) a good feature selection is essential for text classification to make it tractable for machine learning, and to improve classification performance. It should improve classification accuracy and conserve computation, storage, network resources needed for training, and all future use of the classifier. Conversely, poor feature selection limits performance, and no degree of clever induction can make up for a lack of predictive signal in the input features. Rogati and Yang (2002) state that aggressive reduction of the feature space has been repeatedly shown to lead to little accuracy loss and to a performance gain in many cases. Feature selection is important either for improving accuracy or for reducing the complexity of the final classifier. (Keerthi, 2005; Montanes et al, 2004; Montanes et al, 2003; Zheng et al, 2004; Moyotl-Hernandez and Jimenez-Salazar, 2005)

One can refer to the following papers to study more about the importance of feature selection in text categorization task. (Koller and Sahami, 1996; Dash and Liu 1997; Guyon and Elisseeff, 2003; Anghelescu and Muchnik, 2003; Seo et al., 2004; How and Narayanan, 2004; Doan and Horiguchi, 2004b; Yan et al., 2005)

4.3.3 Feature Selection Approaches & Terminologies

In this section, we briefly overview the different concepts and terminologies used in feature selection problems. We find it necessary to identify these concepts as they form the basis of feature selection structure. In the following sections, first the difference between supervised and unsupervised feature selection would be stated. Then two common approached in supervised feature selections would be explained and compared. And finally the distinction between the local and global feature selection would be made.

4.3.3.1 Supervised vs. Unsupervised Feature Selection

Feature selection methods have been successfully applied to text categorization but seldom applied to text clustering due to the unavailability of class label information. (Liu et al., 2003) They define text clustering as one of the central problems in text mining and information retrieval area which group similar documents together.

The performance of clustering algorithms will decline dramatically due to the problems of high dimensionality and data sparseness (Agrawal and Yu, 2000 cited by Liu et al., 2003). Therefore it is highly desirable to reduce the feature space dimensionality.

Depending on if the class label information is required, feature selection can be either unsupervised or supervised. (Liu et al., 2003, 2005) In text categorization the class label information is unavailable hence the dimensionality reduction in text clustering is referred to as unsupervised feature selection while in classification task it is called supervised feature selection.

Any traditional feature selection method that does not need the class information, such as document frequency (DF) and term strength (TS) can be easily applied to clustering. (Law et al., 2002 cited by Liu et al., 2003) There are also other new methods suggested for unsupervised feature selection. Entropy-based feature ranking method (En) is proposed by Dash and Liu (2000) in which the term is measured by the entropy reduction when it is removed. Another method proposed is called Term Contribution (TC) which can be found in Liu et al. (2003). A new proposed method called Term Variance is introduced by Liu et al. (2005). As the focus of this study is on classification task, the supervised feature selection would be the primary concern and unsupervised feature selection is not discussed in detail. The supervised feature selection would be reviewed thoroughly in section 4.3.4

4.3.3.2 Filter Approach vs. Wrapper Approach

There are two main approaches to the problem of feature selection for supervised learning. The filter approach (John et al., 1994) and the wrapper approach (Kohavi and John, 1997). The filter approach scores features independently of the classifier, while the wrapper approach jointly computes the classifier and the subset of features. While the wrapper approach is arguably the optimum approach, for applications such as text classification where the number of features ranges from dozens to hundreds of thousands it can be prohibitively expensive. (Boulis and Ostendorf, 2005)

Langley (1994) grouped different feature selection methods based on their dependence on the inductive algorithm that will finally use the selected subset. *Filter* methods are independent of the inductive algorithm, whereas *wrapper* methods use the inductive algorithm as the evaluation function. See Figure 4.1 and 4.2 for better understanding of these two approaches.

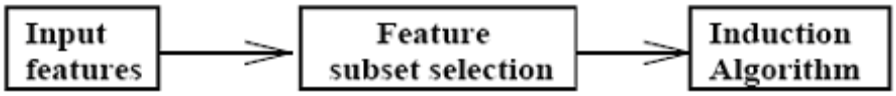


Figure 4.1: The Feature Filter Model

Source: John et al., 1994

As stated by Koller and Sahami (1996) in the filter model, feature selection is performed as a preprocessing step to induction. Thus the bias of learning algorithm does not interact with the bias inherent in the feature selection algorithm. Whereas in the wrapper model the feature selection is being "wrapped around" an induction algorithm, so that the bias of the operators strongly interacts. According to Doan and Horiguchi (2004b) the wrapper method is relatively difficult to implement, especially with a large amount of data. Instead, the filtering approach is usually chosen because it is easily understood for its independent classifiers. In text categorization, the filter approach is often used.

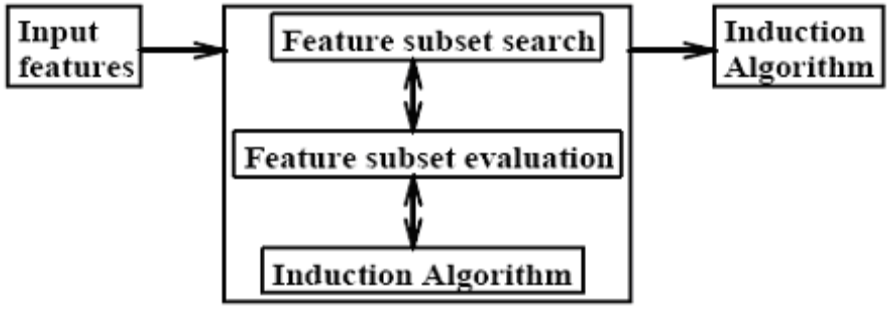


Figure 4.2: The Wrapper Model

Source: John et al., 1994

Guyon and Elisseeff (2003) explain some of the arguments about these two different approaches. They state that compared to wrappers, filters are faster while some filters (e.g. those based on mutual information criteria) provide a generic selection of variables, not tuned for/by a given learning machine. Another compelling justification is that filtering can be used as a preprocessing step to reduce space dimensionality and overcome over fitting. They have also introduced a third approach called embedded which combines the two approaches into one by embedding a filter feature selection method into the process of classifier training. There are other researchers who have mentioned the differences between the filter and the wrapper approach. (Blum and Langley, 1997; Mladenic, 1998; Novovicova and Malik, 2005; Wang Y., and Wang X.J., 2005)

4.3.3.3 Local vs. Global Feature Selection

There are two distinct ways of viewing feature selection, depending on whether the task is performed locally or globally. In local feature selection, for each category, a set of terms is chosen for classification based on the relevant and irrelevant documents in that category. In global feature selection, a set of terms is chosen for the classification under all categories based on the relevant documents in the categories. (Zheng and Srihari, 2003; Novovicova and Malik, 2005; How and Narayanan, 2004)

Preview research on text categorization (Apte et al., 1994) suggests two possible ways in which the word to be used as features can originate from the relevant texts only (local dictionary), or from both the relevant and irrelevant texts (universal dictionary). Apte et al. (1994) and Ng et al. (1997) reported results indicate that local dictionary gives better performance and better accuracy for feature selection problems.

Bong et al. (2005) explain in details the difference between the local and global feature selection (how to select features) and local and universal dictionary (what features to select). In their paper called “An Examination of Feature Selection Framework in Text Categorization” they not only explain the local and global feature selection but also investigate the correspondence effect of local and global feature selection on different

dictionary namely local and universal dictionary. They have proved that local feature selection always outperforms global feature selection in both local and global dictionary and have concluded that local feature selection with local dictionary is the best performer among all. Universal dictionary in either local or global feature selection does not perform as well as the local dictionary. Local dictionary is the better descriptor for the category, and incorporating negative feature does not always optimize the categorization results as claimed by Zheng et al. (2004) and Zheng and Srihari (2003). For the exact details of these approaches and their comparisons refer to the mentioned article.

4.3.4 Feature Selection Metrics in Supervised Filter Approach

FS is commonly performed in TC by keeping the words with highest score according to a measure. They consider the distribution of the words over the different categories. (Montanes et al., 2004) A number of researchers have recently addressed the issue of feature subset selection. There exists several feature selection measures which have been explored in the text filtering literature. These measures include Document Frequency (DF), Information Gain (IG), Mutual Information (MI), Chi-square (CHI), Correlation Coefficient (CC), Odds Ratio (OR), GSS Coefficient (GSS), and Term Strength (TS). (van Rijsbergen, 1979; Quinlan, 1986; John et al., 1994; Lewis and Ringuette, 1994; Schutze et al., 1995; Mitchell, 1996; Ng et al., 1997; Yang and Pedersen, 1997; Mladenic, 1998; Galavotti et al., 2000; Sebastiani, 1999, 2002; Chakrabarti, 2003; Zheng et al., 2003; Forman, 2003)

Many novel FS approaches, such as PCA based algorithm (Malhi and Gao, 2004 cited by Yan et al., 2005), Margin based algorithm (Gilad-Bachrach et al., 2004, cited by Yan et al., 2005), and SVM-based algorithm (Hardin et al., 2004, cited by Yan et al., 2005) were proposed in the past decades. According to Wang et al. (2006) in the text domain, the most popular used FS algorithms are still the traditional ones such as Information Gain, Chi-Square Test, Document Frequency, and Mutual Information. Actually there are other metrics that have been proposed so far, but for the sake of time and space we only mention the most important ones. In this section we briefly give the definition and formulas of the core feature selection metrics from the three main sources

of comparison studies. (Yang and Pedersen, 1997; Forman, 2002, 2003; Prabowo and Thelwall, 2006) The focus of this section relies on the evaluation and comparison of these metrics. The full details of the metric which is going to be used in this study would be explained in Chapter 5.

Document frequency (DF): Document frequency is the number of documents in which a term occurs. The document frequency for each unique term in the training corpus is computed and it removes from the feature space those terms whose document frequency are less than some predetermined threshold. This method is used in Apte et al. (1994). DF thresholding is the simplest technique for vocabulary reduction. However, it is usually considered an ad hoc approach to improve efficiency, not a principled criterion for selecting predictive features.

Information Gain (IG): Information gain is frequently employed as a term goodness criterion in the field of machine learning. It measures the number of bits of information obtained for category prediction by knowing the presence or absence of a term in a document. For each unique term the information gain is computed, and features of feature space whose gain is less than some predetermined threshold will be removed. Lewis and Ringuette (1994) used information gain measure to aggressively reduce the document vocabulary in a Naïve Bayes model and a Decision-Tree approach to binary classification.

Mutual information (MI): Mutual information is a criterion commonly used in statistical language modeling of word associations and related applications. Wiener et al. (1995) used mutual information and Chi-square statistic to select features for input to neural networks. It has also been used by Battiti (1994), Baker and MaCallum (1998), Sebastiani (2002) and Yang (1999). A weakness of mutual information is that the score is strongly influenced by the marginal probabilities of terms. For terms with an equal conditional probability Pr(t|c), rare terms will have a higher score than common terms. The scores, therefore, are not comparable across terms of widely differing frequency.

Chi square statistic (CHI): The chi square statistic measures the lack of independence between t and c and can be compared to the chi square distribution with one degree of freedom to judge extremeness. The chi square statistic has a natural value of zero if t and c are independent. For each category the chi square statistic between each unique term in a training corpus and that category is computed. A major difference between CHI and MI is that chi square is a normalized value, and hence its values are comparable across terms for the same category. However, this normalization breaks down if any cell in the contingency table is lightly populated, which is the case for low frequency terms. Hence, the chi square statistic is known not to be reliable for low-frequency terms. (Dunning, 1993)

Term strength (TS): Term strength is originally proposed and evaluated by Wilbur and Sirotkin (1992) for vocabulary reduction in text retrieval, and later applied by Yang (1995) to reduce the noise for computational efficiency improvements in a statistical learning method for text categorization using LLSF method. Yang and Wilbur (1996) also used term strength to reduce the variables in linear regression and nearest neighbor classification. This method estimates term importance based on how commonly a term is likely to appear in "closely-related" documents. It uses a training set of documents to derive document pairs whose similarity is above a threshold. This method is not task-specific, means that it does not use information about term-category associations.

Odds ratio (OR) Odds ratio was proposed originally by Rijsbergen (1979) for selecting terms for relevance feedback. The basic idea is that the distribution of features on the relevant documents is different from the distribution of features on the non-relevant documents. It has been used by Mladenic (1998) for selecting terms in text categorization.

Correlation Coefficient (CC): Correlation coefficient was defined by Ng et al. (1997) and Sebastiani (2002). It is a variant of the CHI metric, where $CC^2 = \chi^2$. **GSS Coefficient (GSS):** GSS coefficient is another simplified variant of the chi square statistic proposed by Galavotti et al. (2000)

The main feature selection metrics applicable in text categorization and their mathematical formula is brought in Table 4.1. The notations used in table include t_k which indicates the k^{th} term and c_i which is the i^{th} category.

Table 4.1: The Core Metrics in Text Feature Selection and Their Mathematical Form

Source: Sebastiani, 1999 and 2002

METRIC	DENOTED	MATHEMATICAL FORM
Document frequency	$\#(t_k, c_i)$	$P(t_k, c_i)$
Information Gain	$IG(t_k, c_i)$	$P(t_k, c_i) \cdot log\frac{P(t_k, c_i)}{P(c_i) \cdot P(t_k)} + P(\bar{t}_k, c_i) \cdot log\frac{P(\bar{t}_k, c_i)}{P(c_i) \cdot P(\bar{t}_k)}$
Mutual information	$MI(t_k, c_i)$	$\log \frac{P(t_k, c_i)}{P(t_k) \cdot P(c_i)}$
Chi-Square	$\chi^2(t_k, c_i)$	$\frac{N[P(t_k, c_i) \cdot P(\bar{t}_k, \bar{c}_i) - P(t_k, \bar{c}_i) \cdot P(\bar{t}_k, c_i)]^2}{P(t_k) \cdot P(\bar{t}_k) \cdot P(c_i) \cdot P(\bar{c}_i)}$
Corelated Coefficient	$CC(t_k, c_i)$	$\frac{\sqrt{N}[P(t_k, c_i) \cdot P(\bar{t}_k, \bar{c}_i) - P(t_k, \bar{c}_i) \cdot P(\bar{t}_k, c_i)]}{\sqrt{P(t_k) \cdot P(\bar{t}_k) \cdot P(c_i) \cdot P(\bar{c}_i)}}$
Term strength	$s(t)$	$P_r(t \in y \mid t \in x)$
Odd Ratio	$OR(t_k, c_i)$	$\frac{P(t_k \mid c_i) \cdot [1 - P(t_k \mid \bar{c}_i)]}{[1 - P(t_k \mid c_i)] \cdot P(t_k \mid \bar{c}_i)}$
GSS Coefficient	$GSS(t_k, c_i)$	$P(t_k, c_i) \cdot P(\bar{t}_k, \bar{c}_i) - P(t_k, \bar{c}_i) \cdot P(\bar{t}_k, c_i)$

One of the important issues in text categorization problems is the choice of the feature selection metrics. We are going to review feature selection methods to find out which of them are both computationally scalable and high performing across different classifiers and data collections. In the following section we review the related works regarding the comparison of feature selection metrics in text categorization task.

Evaluation & Comparisons of Feature Selection Metrics

Yang and Pedersen (1997) have made a comparative study on feature selection methods in statistical learning of text categorization. They compare five term selection scores based on 3 criteria which relates to the nature of the metrics themselves, including their favoring toward common terms, their usage of category information (being task-sensitive or task-free), and their usage of term absence. (Refer to Table 4.2) They performed their experiments on two datasets, Reuters-22173 and Ohsumed, and under

two classifiers, kNN and a Linear Least Square Fit (LLSF). Yang and Pedersen (1997), supported by Sebastiani's (2002) automatic text categorization review article, suggest that IG and χ^2 have similar performance in supporting document classification and aggressive term removal, with both being significantly better than MI. Of these, χ^2 seems to be gaining support in practice for classification, perhaps for its calculation efficiency (Ng et al. 1997) although selecting terms by using χ^2 has also been criticized for being unreliable when the cell value is less than 5 (Dunning, 1993). DF thresholding is found comparable to the performance of IG and CHI with up to 90% term removal, while TS is comparable with up to 50-60% term removal. Mutual information has inferior performance compared to the other methods due to a bias favoring rare terms and a strong sensitivity to probability estimation errors.

Table 4.2: Criteria and Performance of Feature Selection Methods in kNN and LLSF
Source: Yang and Pedersen, 1997

Method	DF	IG	CHI	MI	TS
favoring common terms	Y	Y	Y	N	Y/N
using categories	N	Y	Y	Y	N
using term absence	N	Y	Y	N	N
performance in kNN/LLSF	excellent	excellent	excellent	poor	ok

Forman (2003) presents an empirical study of twelve feature selection metrics evaluated on a benchmark of 229 text classification problem instances that originated from Reuters, Ohsumed, TREC and etc. He used Support Vector Machine (SVM) as a classifier as Yang and Pedersen (1997) had not considered SVM, which they later found to be superior to the algorithms they had studied, LLSF, and kNN. The results on these benchmark datasets showed that the well-known Information Gain metric was not best for the goals of F-measure, Recall, or Accuracy, but instead an outstanding new feature selection metric, Bi-Normal Separation. For the goal of Precision alone, however, Information Gain was superior. Forman (2003) provided evidence that IG and CHI have correlated failure, hence choosing optimal pairs of scores, these two works poorly together.

Rogati and Yang (2002) examined major feature selection criteria (IG, DF, CHI, IG2) using two benchmark collections; Reuters 21578 and part of RCV1. Four well-known classification algorithms were used including a Naive Bayesian (NB) approach, a Rocchio-Style classifier, a k-Nearest Neighbor (kNN) method, and a Support Vector Machine (SVM) system. The empirical results (Figure 4.3) of their study suggest filter methods which include the $\chi 2$ statistic, combining them with DF or IG outperforms those based on other criteria. SVM was the least sensitive to different feature selection methods.

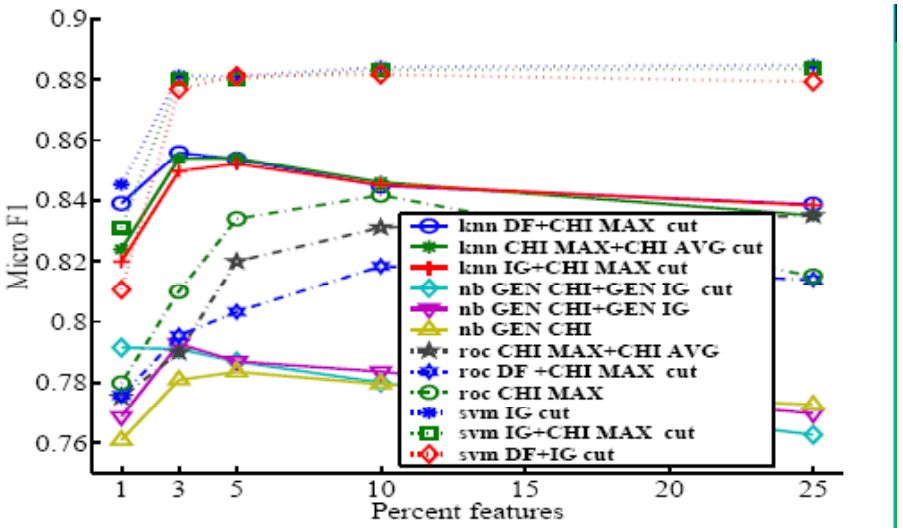


Figure 4.3: Top Three Feature Selection Methods for Reuters 21578 (Micro F1)

Source: Rogati and Yang, 2002

Eyheramendy and Madigan (2005) introduced a new feature selection method called PIP that evaluates the posterior probability of inclusion of a given feature over all possible models. They compare their new feature selection scores with five other feature selection scores that have been prominent in the previous studies including IG, BNS, CHI, OR and Word Frequency. They evaluated these feature selection scores on two widely used benchmark text classification datasets, Reuters-21578, and 20-Newsgroup, with 4 classification algorithm namely Multinomial, Poisson, and Binary Naïve Bayes and the Hierarchical Probit Classifier. They state that their results are consistent with Yang and Pedersen (1997) in that CHI and IG seem to be the strong scores in most instances. They resulted that CHI, BNS, and PIP are the best performing scores. But this should be considered that feature selection scores and classification algorithms seem to be highly

data and model dependent. It means that feature selection scores act differently on different datasets and classification algorithm.

Seo et al. (2004) examined four feature selection methods in their study in order to extract a set of candidate words for concept words in ontology learning. They considered MI, CHI, Markov Blanket, and IG. Information gain had the worst performance in identifying word features in the ontology. The Markov Blanket feature selection method performed better than information gain. The mutual information and CHI methods were far superior to the other two feature selection methods. Between the two, there is no clear winner. They were able to identify words with high semantic content for the class, as required for ontology.

Prabowo and Thelwall (2006) made a comparison of feature selection methods for RSS (Rich Site Syndication) feed corpus. They compare CHI, MI, and IG. According to the evaluation used, χ^2 seems to be the best method to determine term significance for RSS feeds. The miss rates indicate that MI and IG are less effective than χ^2 for the aggressive exclusion of terms. In addition, χ^2 tends to agree with both MI and IG, with MI and IG showing significant disagreement. The χ^2 method, however, is far from perfect as an extremely high value can be assigned to relatively insignificant terms; hence the higher the degree of the significance of a term, should be treated cautiously.

Mladenic and Grobelnik (1999) found that OR is the strongest feature selection score. They compared 11 scores and performed their experiments on a Naïve Bayes model using the Yahoo dataset.

Swan and Allan (1999, 2000) have adopted χ^2 for extracting time varying features from text. CHI has been proved to be an effective and robust feature selection measure in the literature. (Zheng and Srihari, 2003) Both IG and CHI are dominant in the area of text categorization since they have been proved to be very effective and efficient. (Yan et al. 2005)

Fragos et al. (2005) used maximum entropy for text classification, using a feature selection strategy and assigning weights to the features with the chi-square test. The

result of the evaluation was very promising. Schutze et al. (1995) also have used chi square metric as the feature selection method for their experiment.

Despite previous research into term selection methods, there is no clear indication of the superiority of any particular method for all types of data as each has its own strengths and weaknesses. There is no strong evidence, however, to suggest which method is the most effective as they are highly dependent on the datasets and the classification algorithms. A number of feature selection metrics have been explored in text categorization, among which information gain (IG), chi-square (CHI), correlation coefficient (CC) and odds ratios (OR) are considered most effective. An overview on the comparative study of different feature selection metrics has proved that the classical feature selection scores are still the best ones for text categorization task. A thorough comparative experiment performed across different classifier induction methods and different document corpora have shown chi square to be one of the most effective feature selection methods especially for time-varying features, allowing to reduce dimensionality of the feature space with no or small loss of effectiveness. The purpose of this study is to extract the time varying features from news stories to predict the stock price movement. According to the literature above, we think that the chi-square metric would be the best feature selection method to be used in our study.

4.4 Document Representation

Document representation is the final task in document preprocessing. The documents are represented in terms of those features to which the dictionary was reduced in the precedent steps. Thus, the representation of a document is a feature vector of n elements where n is the number of features remaining after finishing the selection process.

A major challenge of the text classification problem is the representation of a document. Document indexing refers to the task of automatically constructing internal representations of the documents that be amenable to interpretation by the classifier induction algorithm and compactly capture the meaning of the documents. (Galavotti et

al., 2000) All the algorithms applied to TC need the documents to be represented in a way suitable for the induction of the classifier.

When choosing a document representation, the goal is to choose the features that allow document vectors belonging to different categories to occupy compact and disjoint regions in the feature space (Jain et al., 2000). There exist different types of information that can be extracted from documents for representation. The simplest is the Bag-of-Words representation (BOW) in which each unique word in the training corpus is used as a term in the feature vector. Second type is the categorized proper names and named entities (CAT) that only uses the tokens identified as proper names or named entities from the training corpus used for representation. (Yilmazel et al., 2005)

A recent comprehensive study (Moschitti and Basili, 2004) surveys the different approaches in document representation that have been taken thus far and evaluates them in standard text classification resources. The conclusion implies that more complex features do not offer any gain when combined with state-of-the-art learning methods, such as Support Vector Machines (SVM). (Boulis and Ostendorf, 2005)

4.4.1 Vector Space Model

The simplest and almost universally used approach is the bag-of-word representation, where the document is represented with a vector of the word counts that appear in it. It consists of identifying each document with a numerical vector whose components weight the importance of the different words in the document. (Salton and McGill, 1983; Salton, 1989) According to Chakrabarti (2003) in the vector-space model, documents are represented as vectors in a multidimensional Euclidean space. Each axis in this space corresponds to a term (token). Despite the simplicity of such a representation, classification methods that use the bag-of-words feature space often achieve high performance. (Boulis and Ostendorf, 2005)

4.4.2 Term Weighting Methods in Vector Space Modeling

Term Weighting is one of the important issues in text categorization. Originally term weighting has been widely investigated in information retrieval. According to How and Narayanan (2004) term weighting has been used to weight representative term used to describe and summarize document content based on a term's importance. The coordinate of document *d* in the direction corresponding to term *t* is determined by two quantities: Term Frequency (TF) and Inverse Document Frequency (IDF). (Chakrabarti, 2003)

There are three assumptions that, in one form or another, appear in practically all weighting methods: 1. "rare terms are no less important than frequent terms" which is called the IDF assumption; 2. "multiple appearances of a term in a document are no less important than single appearances" which is a TF assumption; 3. "for the same quantity of term matching, long documents are no more important than short documents" which is called this the normalization assumption.

Term Frequency

It is the simplest measure to weight each term in a text. In this method, each term is assumed to have importance proportional to the number of times it occurs in a text. The weight of a term t in a text d is given by:

$$tf(t_k, d_j) = \begin{cases} 1 + \log \#(t_k, d_j) & \text{if } \#(t_k, d_j) > 0 \\ 0 & \text{otherwise} \end{cases}$$

Where # (t_k ,d_j) denotes the number of times t_k occurs in d_j. Term frequency is known to improve recall in information retrieval, but does not always improve precision. Because frequent terms tend to appear in many texts, such term have little discriminative power. In order to remedy this problem, terms with high frequency are usually removed from the term set. Finding optimal thresholds is the main concern in this method. (Tokunaga and Iwayama, 1994)

Inverse Document Frequency

While term frequency concerns term occurrences within a text, inverse document frequency (IDF) concerns term occurrence across a collection of texts. The intuitive meaning of IDF is that terms which rarely occur over a collection of texts are valuable. The importance of each term is assumed to be inversely proportional to the number of texts that contain the term. The IDF factor of a term t is given by:

$$\log \frac{|Tr|}{\#_{Tr}(t_k)}$$

Where Tr is the total number of texts in the collection and denominator is the number of text that contains the term t_k. Since IDF represents term specificity, it is expected to improve the precision. (Tokunaga and Iwayama, 1994)

Term Frequency Inverse Document Frequency

Salton and Yang (1973) proposed to combine term frequency and inverse document frequency to weight terms, and showed that the product of them gave better performance. The combination weight of a term t in a text is given by:

$$tfidf(t_k, d_j) = tf(t_k, d_j) \cdot \log \frac{|Tr|}{\#_{Tr}(t_k)}$$

Term Frequency Inverse Document Frequency known as *tfidf* is one of the most popular term weighting schemes in information retrieval. *tfidf* assumes that "multiple appearances of a term in a document are more important than single appearances" and "rare terms are more important than frequent terms". It has gained popularity in text categorization assuming that the index terms are mutually independent. (How and Narayanan, 2004) Weights obtained by *tfidf* Equation are usually normalized to unit length by cosine normalization, which enforce the normalization process.

$$w_{kj} = \frac{tfidf(t_k, d_j)}{\sqrt{\sum_{s=1}^{|\mathcal{T}|} tfidf(t_s, d_j)^2}}$$

4.5 Classifier Learning

After the phase of document representation, it would be the time for classifier learning, in which a classifier is generated from weighted representations of the training documents. In this phase, an automated learning algorithm learns the necessary association knowledge from the training examples to build a classifier for each category. (Ng et al., 1997)

Until the late ’80s the most popular approach to TC, at least in the “operational” community, was a *knowledge engineering* (KE) one, consisting in manually defining a set of rules encoding expert knowledge on how to classify documents under the given categories. In the ’90s this approach has increasingly lost popularity (especially in the research community) in favor of the *Machine Learning* (ML) paradigm, according to which a general inductive process automatically builds an automatic text classifier by learning, from a set of pre-classified documents, the characteristics of the categories of interest. The advantages of this approach are accuracy comparable to that achieved by human experts, and a considerable savings in terms of expert labor power, since no intervention from either knowledge engineers or domain experts is needed for the construction of the classifier or for its porting to a different set of categories. (Sebastiani, 2002)

A growing number of statistical machine learning techniques have been applied to text categorization in recent years, and a number of researchers have proposed various techniques to carry out an automatic classification task, including: Hierarchic Classification (Jenkins et al., 1999; Larkey, 1998), Multivariate Regression Models (Fuhr et al., 1991; Yang and Chute, 1994; Schutze et al., 1995), K-Nearest Neighbor Classification (Creecy et al.,1992; Yang, 1994,1999; Kwon and Lee, 2000), Bayesian Classification (Tzeras and Hartman, 1993; Lewis and Ringuette, 1994; McCallum and Nigam, 1998; Lam et al., 1997), Decision Tree (Lewis and Ringuette, 1994), Neural Network Based Classification (Chen et al., 1994; Yang, 1994; Wiener et al., 1995; Ng et al., 1997; Ruiz and Sririvasan, 1999), Rocchio Algorithm (Buckley et al., 1994, Joachims, 1997) Symbolic Rule Learning (Apte et al., 1994; Cohen and Singer, 1996). A number of

linear classification methods such as the Linear Least Squares Fit (LLSF), logistic regression, and Support Vector Machines Based Classification (Cortes and Vapnik, 1995; Joachims, 1998, 2002; Kowk, 1998; Dumais et al., 1998; Hearst et al., 1998; Dumais and Chen, 2000; Siolas and d'Alche-Buc, 2000) have been applied to text categorization problems. These methods share the similarity by finding hyperplanes that approximately separate a class of document vectors from its complement. However, support vector machines are so far considered special in that they have been demonstrated to achieve the state of the art performance. (Zhang and Oles, 2001)

4.5.1 Comparison of Categorization Methods

As stated above, many algorithms have been proposed for document categorization. Some papers compare the effectiveness of selected algorithms. We provide a brief overview in this domain.

As stated by Brucher et al. (2002) the results may be heavily dependent of the test data set. Several parameters usually have to be defined to initialize the procedures and the performance may depend on their initialization. If various "standard data sets" exist, the algorithms may even be tuned to deliver high efficiency for these data sets. Taken these limitations into consideration, we have to emphasize that the SVM method has outperformed the other methods in several comparisons.

Dumais et al. (1998) compare 5 different learning methods including Find Similar (a variant of Rocchio's method for relevance feedback), Decision Trees, Naïve Bayes, Bayes Net, and Support Vector Machines. They use Reuter-21578 collection as their datasets. The accuracy of their simple linear SVM is among the best reported for the Reuters-21578 collection. According to them, Linear Support Vector Machines (SVMs) are particularly promising because they are very accurate, quick to train, and quick to evaluate. Figure 4.4 represent the result of their experiment.

Joachims (1998) explores the use of Support Vector Machines (SVMs) for learning text classifiers and compares this method with 4 other conventional learning methods namely Naïve Bayes classifier, Rocchio algorithm, K-nearest neighbor classifier,

and C4.5 decision tree/rule learner. Test collections include both Reuter-21578 and Ohsumed. The experimental results show that SVMs consistently achieve good performance on text categorization tasks, outperforming existing methods substantially, and significantly. They also found that SVM using RBF kernels perform better than those with polynomial kernels.

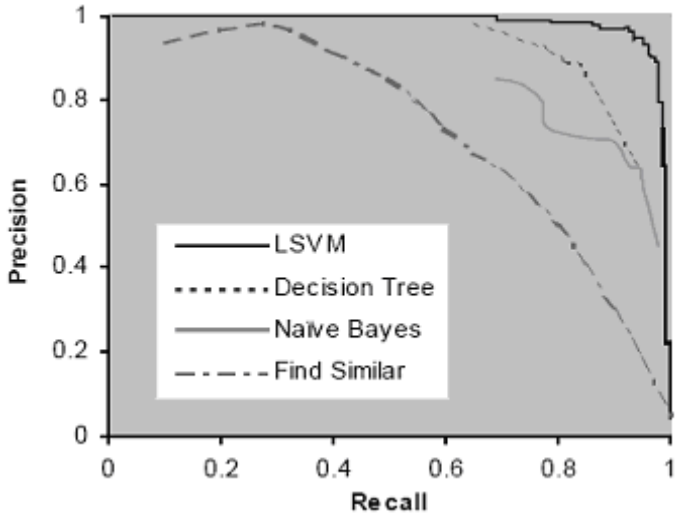


Figure 4.4: Comparison of Text Classifiers

Source: Dumais et al., 1998

Yang and Liu (1999) make a comprehensive comparison of five learning algorithms including SVM, kNN, LLSF (Linear Least Squares Fit), Naïve Bayes, and Neural Network. For the micro-level performance on pooled category assignments, SVM and kNN significantly outperform the other classifiers, while NB significantly underperforms all the other classifiers. With respect to the macro-level (category-level) performance analysis, all the significance tests suggest that SVM, kNN, and LLSF belong to the same class, significantly outperforming NB and NNet.

Siolas and d'Alche-Buc (2000) make a comparison between SVM and KNN in their experiment, which is tested on the 20-Newsgroups database. Support Vector Machines perform better than kNN and provide the best accuracy on test data. The SVM results in terms of precision and recall appear to be very high.

Basu et al. (2003) compares an artificial neural net algorithm with a support vector machine algorithm for use as text classifiers of news items from Reuters-21578. In the overall comparison of SVM and ANN algorithms for this data set, the results over all conditions for both recall and precision indicate significantly differences in the performance of the SVM algorithm over the ANN algorithm and of the reduced feature set over the larger feature set. Recognizing that SVM is a less (computationally) complex algorithm than the ANN, they concluded that SVM is preferable at least for this genre of data, i.e., many short text documents in a relatively few well populated categories.

SVMs have shown to yield good generalization performance on a wide variety of classification problems, most recently text categorization. The empirical evidence has proven this matter. There exists some theoretical analysis which explains why SVM perform well for text categorization. (Joachims, 1998) The following arguments give theoretical evidences that SVMs should perform well for text categorization.

High dimensional input space: When learning text classifiers, one has to deal with very many features. Since SVMs use over fitting protection, they have the potential to handle these large feature spaces. **Document vectors are sparse:** For each document, the corresponding document vector contains only few entries which are not zero. SVMs are well suited for problems with dense concepts and sparse instances. **Most text categorization problems are linearly separable:** All Ohsumed and many of the Reuters categories are linearly separable, and the idea of SVMs is to find such linear (polynomial or RBF, etc.) separators.

SVMs have proved both empirically and theoretically to be well suited for text categorization. (Joachims, 1998) One of the advantages of SVMs over the conventional methods is their robustness. SVMs show good performance in all experiments, avoiding catastrophic failure, as observed with the conventional methods on some tasks. Furthermore, SVMs do not require any parameter tuning, since they can find good parameter settings automatically. All this makes SVMs a very promising and easy-to-use method for learning text classifiers from examples. In the following section, we explain this method and its performance on text categorization problems.

4.5.2. Support Vector Machines (SVMs)

Support Vector Machine (SVM) (Cortes and Vapnik, 1995; Vapnik, 1998) is a state-of-the-art classification algorithm that is shown to be particularly successful in text categorization (Joachims, 1998; Dumais et al., 1998). It is a machine learning method for solving two-class pattern recognition problems and has been shown to yield good generalization performance on a wide variety of classification problems that require large-scale input space, such as handwritten character recognition, face detection and text categorization. (Dumais et al., 1998; Joachims, 1998; Yang and Liu, 1999; Fukumoto and Suzuki, 2001) SVM is based on statistical learning theory, which uses the principle of Structural Risk Minimization (SRM) instead of Empirical Risk Minimization which is commonly employed with other statistical methods. SRM minimizes the upper bound on the generalization error, as against ERM which minimizes the error on the training data. Thus, SVMs are known to generalize better. (Tay et al., 2003)

The idea of SRM is to find a hypothesis *h* which reflects the well-known trade-off between the training error and the complexity of the space. SVM learns from the training set to find a decision surface (classifier) in the vector space of documents that best separates the data points (documents) into two classes (relevant and non-relevant). The decision surface by SVM for linearly separable space is a hyperplane which can be written as the following equation where x is an arbitrary feature vector and w and b are learned from a training set of linearly separable data. (Masuyama and Nakagawa, 2002)

$$\mathbf{w} \cdot \mathbf{x} + b = 0$$

Figure 4.5 illustrates the optimum separation hyperplane. (SVM Portal, 2005) The solid line is the hyperplane and the two dashed lines parallel to the solid line indicate the boundaries in which one can move the solid line without any misclassification. SVM finds the solid line which maximizes the margin (distance) between those parallel dashed lines. The training documents which lie on either of two dashed lines are called support vectors. The purpose of SVM is to find the hyperplane that best classifies the training datasets (documents) with the maximum accuracy rate and minimum error rate.

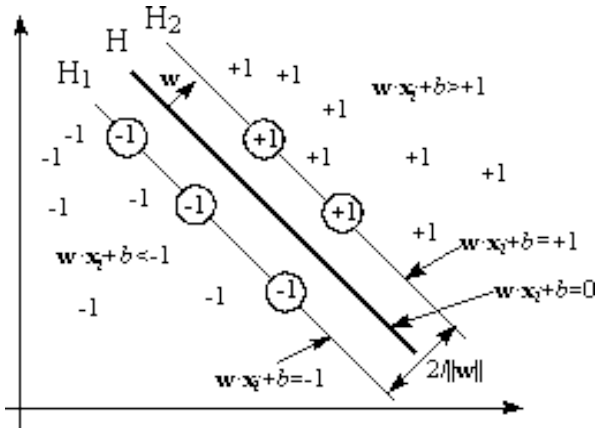


Figure 4.5: The Optimum Separation Hyperplane (OSH)

Source: The "SVM" Portal, 2005

The aim of SVM is to maximize the distance between the two parallel lines which can be expressed as the following equations. The distance between these two lines is equal to 2/||**w**||.

$$\mathbf{x} \cdot \mathbf{w} + b = 1$$

$$\mathbf{x} \cdot \mathbf{w} + b = -1$$

In order to maximize M = 2/||**w**||, we should minimize ||**w**|| under the constraints mentioned above and that can be summarized as the following equation:

$$\begin{aligned} \mathbf{x}_i \cdot \mathbf{w} + b \geq +1 \quad & \text{for } y_i = +1 \\ \mathbf{x}_i \cdot \mathbf{w} + b \leq -1 \quad & \text{for } y_i = -1 \end{aligned}$$

$$y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 \geq 0$$

x_i is a feature vector of the i^{th} training document represented by an n dimensional vector and y_i is the class (positive (+1) or negative (-1)) label of the i^{th} training document. All vectors lying on one side of the hyperplane are labeled as -1, and all vectors lying on the other side are labeled as 1. The training documents which lie on either of two dashed lines are called support vectors.

To solve the problem, it should be switched to Lagrangian formulation. The reason for doing this is that the constraints will be replaced by constraints on the Lagrange multipliers themselves (α_i), which will be much easier to handle. The Lagrangian formulation can be written as the following equation:

$$L_P \equiv \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{l} \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{w} + b) + \sum_{i=1}^{l} \alpha_i$$

Lp should be minimized. This is a convex Quadratic Programming (QP) optimization problem which is very time-consuming. Sequential Minimal Optimization (SMO) is an algorithm for training the SVM where this large QP problem is broken down into a series of smallest possible QP problems which are solved analytically. SMO can handle very large training datasets and considerably speeds up training times. An equivalent dual problem can be solved instead of Lp. L_D is resulted by substituting the Lp with the following equations:

$$L_D = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j$$

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i$$

$$\sum_i \alpha_i y_i = 0$$

Support vector machines use a particular type of function class which are classifiers with large ‘‘margins’’ in a feature space induced by a kernel. A kernel k(x,y) is a similarity measure defined by an implicit mapping ø, from the original space to a vector space (feature space) such that: k(x,y) = ø (x) · ø (y). In the basic form, SVM learns a linear threshold function. Nevertheless, by a simple plug-in of an appropriate kernel function, they can be used to learn polynomial and radial basis function classifiers. (Lee and Kageura, 2006) Two common families of kernels are polynomial kernels and radial basis functions. The choice of the kernel function is crucial to the efficiency of support

vector machines and it is shown that rbf-kernel yields the best performance. (Paaß et al., 2003) The examples of common kernels are as following:

- Linear kernel: $k(x,y) = x^T y$
- Polynomial kernel: $k(x,y) = (\gamma(x^T y) + c)^d$
- Radial basis function (rbf) kernel: $k(x,y) = \exp(- \gamma||x-y||^2)$

4.5.3 Measures of Categorization Effectiveness

Classification effectiveness is measured in terms of the classic IR notations of precision and recall, adapted to the case of document categorization. (Sebastiani, 1999) According to Zheng and Srihari (2003) classification effectiveness has been evaluated in terms of the standard precision, recall, F-measure Rijsbergen (1979) and precision/recall breakeven point. All of these measures can be calculated based on two conventional methods, namely macro-averaging and micro-averaging. (Aas and Eikvil, 1999) Macro-average performance scores are determined by first computing the performance measures per category and then averaging these by the number of categories. Micro-average performance scores are determined by first computing the totals of contingency table for all categories and then use these totals to compute the performance measure of each category. There is an important distinction between the two types of averaging. Micro-averaging gives equal weights to every document, while macro-averaging gives equal weights to each category.

In a binary decision problem, a classifier labels examples as either positive or negative. The decision made by the classifier can be represented in a structure known as a confusion matrix or contingency table. The confusion matrix has four categories: True positives (TP) are examples correctly labeled as positives. False positives (FP) refer to negative examples incorrectly labeled as positive. True negatives (TN) correspond to negatives correctly labeled as negative. Finally, false negatives (FN) refer to positive examples incorrectly labeled as negative. From the contents of this contingency table or confusion matrix, precision and recall can be calculated. (Davis and Goadrich, 2006)

There exist two types of contingency table, a contingency table for each category c_i, and a global contingency table. (Table 4.3 and 4.4) The global contingency table is used for calculating the micro-average measurements. For calculating the macro-average measurement first the micro-average measures should be calculated and then should be divided by the total number of categories. (Sebastiani, 1999)

Table 4.3: The Contingency Table for Category c

Source: Sebastiani, 1999

Category c_i		expert judgments	
		YES	**NO**
classifier judgments	**YES**	TP_i	FP_i
	NO	FN_i	TN_i

Table 4.4: The Global Contingency Table

Source: Sebastiani, 1999

Category set $C = \{c_1, \ldots, c_m\}$		expert judgments	
		YES	**NO**
classifier judgments	**YES**	$TP = \sum_{i=1}^{m} TP_i$	$FP = \sum_{i=1}^{m} FP_i$
	NO	$FN = \sum_{i=1}^{m} FN_i$	$TN = \sum_{i=1}^{m} TN_i$

A common evaluation strategy is to consider classification accuracy or its complement error rate. Accuracy refers to the proportion of correctly assigned documents to the total number of documents assigned and error is the proportion of incorrectly assigned documents to the total number of documents assigned. The better the classifier, the higher would be its accuracy.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

$$\text{Error} = \frac{FP + FN}{TP + FP + TN + FN}$$

The issue of skewed data leads us away from using the standard performance measure of accuracy to evaluate the results. With the positive class so small relative to the negative class, it is trivial to achieve high accuracy by labeling all test examples negative. (Goadrich et al., 2006)

In an imbalanced setting, where the prior probability of the positive class is significantly less than the negative class, accuracy is inadequate as a performance measure since it becomes biased towards the majority class. That is, as the skew increases, accuracy tends towards majority class performance, effectively ignoring the recognition capability with respect to the minority class. In these situations, other performance measures such as precision (in conjunction with recall) may be more appropriate as they remain sensitive to the performance on each class. (Landgrebe et al., 2006) Precision and recall concentrate on the positive examples, since precision measures how accurate we are at predicting the positive class, while recall measures how many of the total positives we are able to identify. (Goadrich et al., 2006)

Precision indicates the number of categories correctly assigned divided by total number of categories assigned and **recall** is the number of categories correctly assigned divided by the total number of categories that should be assigned. (Lewis, 1992) A complete formulation of precision and recall can be summarized in table 4.5.

Table 4.5: The Most Popular Effectiveness Measures in Text Classification
Source: Debole and Sebastiani, 2002

	Precision	Recall
Microaveraging	$\pi = \frac{TP}{TP+FP} = \frac{\sum_{i=1}^{\|\mathcal{C}\|} TP_i}{\sum_{i=1}^{\|\mathcal{C}\|}(TP_i+FP_i)}$	$\rho = \frac{TP}{TP+FN} = \frac{\sum_{i=1}^{\|\mathcal{C}\|} TP_i}{\sum_{i=1}^{\|\mathcal{C}\|}(TP_i+FN_i)}$
Macroaveraging	$\pi = \frac{\sum_{i=1}^{\|\mathcal{C}\|} \pi_i}{\|\mathcal{C}\|} = \frac{\sum_{i=1}^{\|\mathcal{C}\|} \frac{TP_i}{TP_i+FP_i}}{\|\mathcal{C}\|}$	$\rho = \frac{\sum_{i=1}^{\|\mathcal{C}\|} \rho_i}{\|\mathcal{C}\|} = \frac{\sum_{i=1}^{\|\mathcal{C}\|} \frac{TP_i}{TP_i+FN_i}}{\|\mathcal{C}\|}$

An arguably complete view of a system's performance is given by the precision-recall curve, which is commonly summarized in a single indicator using the average precision over various standard recall levels or number of documents. (Goutte and Gaussier, 2005)

Precision-Recall (PR) curves (Figure 4.6) are commonly used to present results for binary decision problems in machine learning (Davis and Goadrich, 2006) and the relationship between the precision and the recall is characterized by the precision-recall curve. (Lee and Chen, 2006) Points are commonly created by first defining a threshold, where all examples ranked higher than this threshold are classified positive and all examples lower than this threshold are classified negative. From these classifications we can calculate our *TP*, *FP*, *TN*, and *FN* values, followed by the true positive rate, false positive rate, recall, and precision for this threshold. The threshold is then varied from the highest rank to the lowest rank, giving us all meaningfully distinct threshold values for these ranking and therefore all possible points on our curve. (Goadrich et al., 2006)

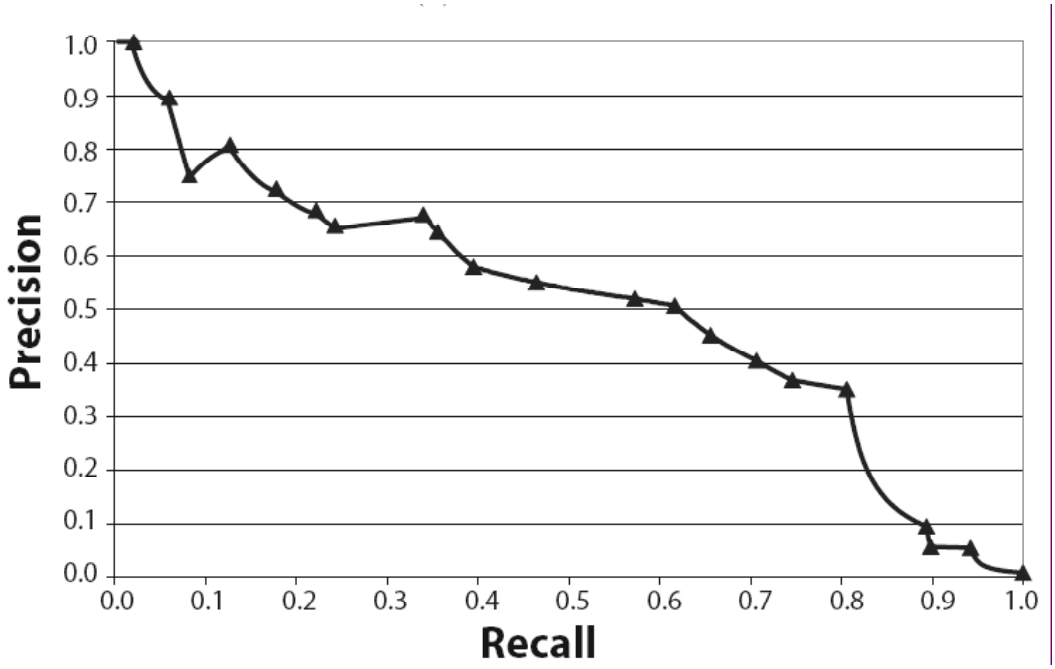


Figure 4.6: Precision-Recall Curve

Source: Goadrich et al., 2006

Precision-Recall break-even point: The performance measures (precision, recall) may be misleading when examined alone. A classifier exhibits a trade-off between recall and precision as obtaining a high recall usually means sacrificing precision and vise versa. If the recall and precision are tuned to have an equal value, then this value is called the break-even point of the system. The break-even point has been commonly used in text categorization evaluations. (Aas and Eikvil, 1999; Fragos et al., 2005)

Precision-Recall F-measure: Another evaluation criterion that combines recall and precision is the F-measure put forwarded by Rijsbergen (1979). It is the weighted harmonic mean of precision and recall which is also known as F1 measure. F1 measure attributes equal importance to precision and recall. The higher the F-measure, the better would be the system performance. For micro-average and macro-average F1, micro-average and macro-average precision and recall would be used respectively.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

4.6 Summary

In this chapter, we reviewed the main processes of text categorization task. It mainly consists of text preprocessing, document representation, induction learning, and performance evaluation. In text preprocessing, the documents are first tokenized, then the words with the same rout are stemmed and the stop words are going to be removed from the text. In document representation, first the feature selection is performed in order to remove the non-informative terms and then the remained terms are weighted according to mentioned criteria. The represented documents are given to the text classifier to be learned and the classification task is then evaluated. In our study we have followed the same procedures. The methods and the techniques that have been used in this study regarding the text categorization task are going to be addressed in Chapter 5.

Chapter 5

Research Methodology

5. Research Methodology

A research design is the strategy for a study and the plan by which the strategy is to be carried out. It specifies the details of how the project should be conducted in order to fulfill the research objectives. Research designs may be broadly classified as Exploratory and Conclusive. In this chapter, we will discuss the research approach of our study and the design strategy implemented to carry out the research question.

5.1 Research Approach and Design Strategy

Exploratory research aims to develop initial hunches or insights and provide direction for any further research needed. The primary purpose of exploratory research is to shed light on the nature of a situation and identify any specific objectives or data needs to be addressed through additional research. Exploratory research is most useful when a decision maker wishes to better understand a situation and/or identify decision alternatives. The objectives of exploration may be accomplished with different techniques. Both qualitative and quantitative techniques are applicable. (Cooper and Schindler, 2003) Qualitative research is an unstructured, primarily design based on small samples which is intended to provide some insight and understanding, but quantitative research seeks to quantify data and typically apply some form of statistical analysis.

Exploration relies more heavily on qualitative techniques, but in some cases, there exist sufficient data to allow data mining or exploration of relationships between individual measurements. Data mining is a useful tool, an approach that combines exploration and discovery with confirmatory analysis which allows decision-makers to be supported through exploratory quantitative research. Primary data are originated by a researcher for the specific purpose of the problem at hand, but secondary data are data that have already been collected for other purposes. Secondary data can be classified as internal and external. (Malhotra and Birks, 2002) Our research strategy is the analysis of secondary data. For conducting our research two types of data are needed: 1. Companies' stock prices which are an internal secondary data gathered from the financial databases and 2. Financial news articles which are gathered from external sources. Research approach and the design strategy of this study is illustrated in Figure 5.1

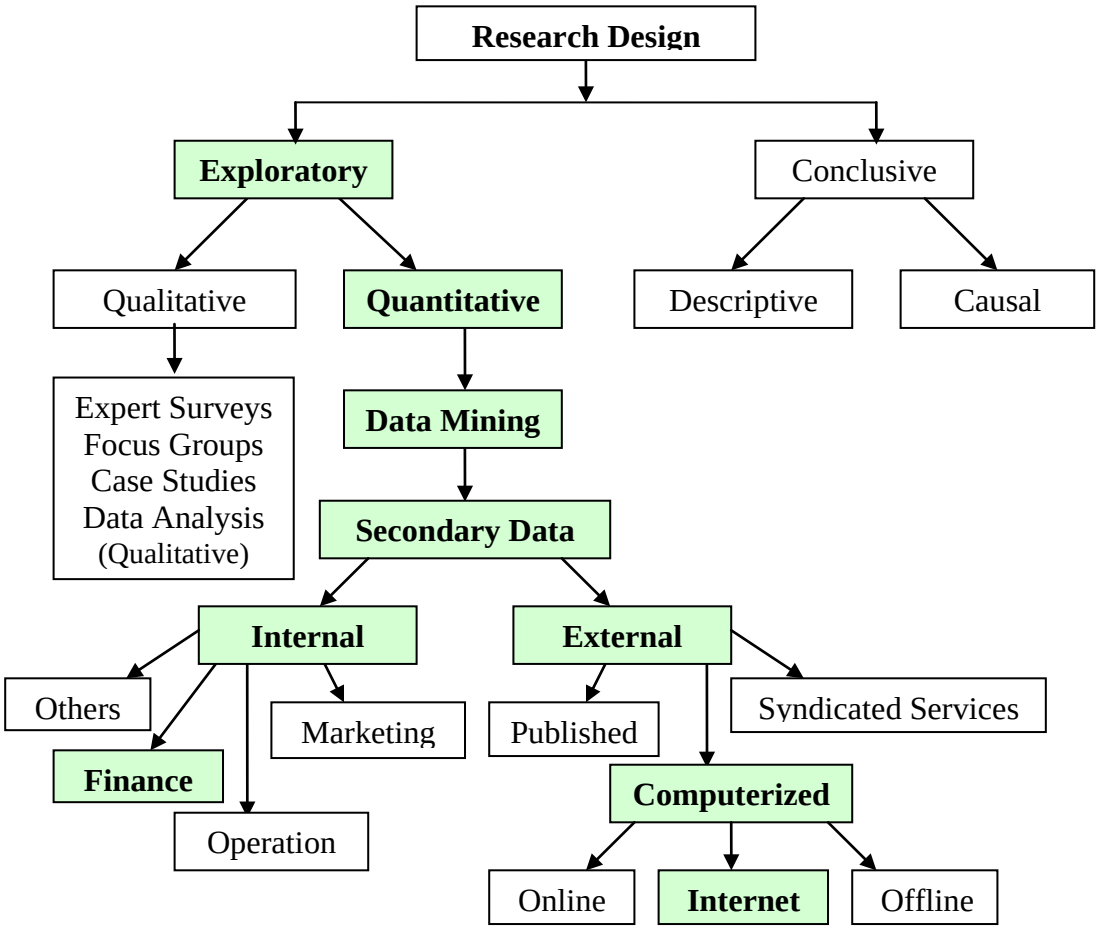


Figure 5.1: Research Approach and Design Strategy of the Study

5.2 The Overall Research Process

In Chapter 2, we reviewed and evaluated different studies related to the prediction of stock price movement using financial news articles. We construct our overall process mostly based on Fung (2005) methodology with some changes and amendments due to the nature of our study. The overall process and the sequence of steps are provided in Figure 5.2. The process consists of many different courses of actions including data collection, data and text preprocessing, feature selection, term weighting, dimension reduction, classifier learning, and system evaluation. The following sections will address the details of the process and the methods associated to each step.

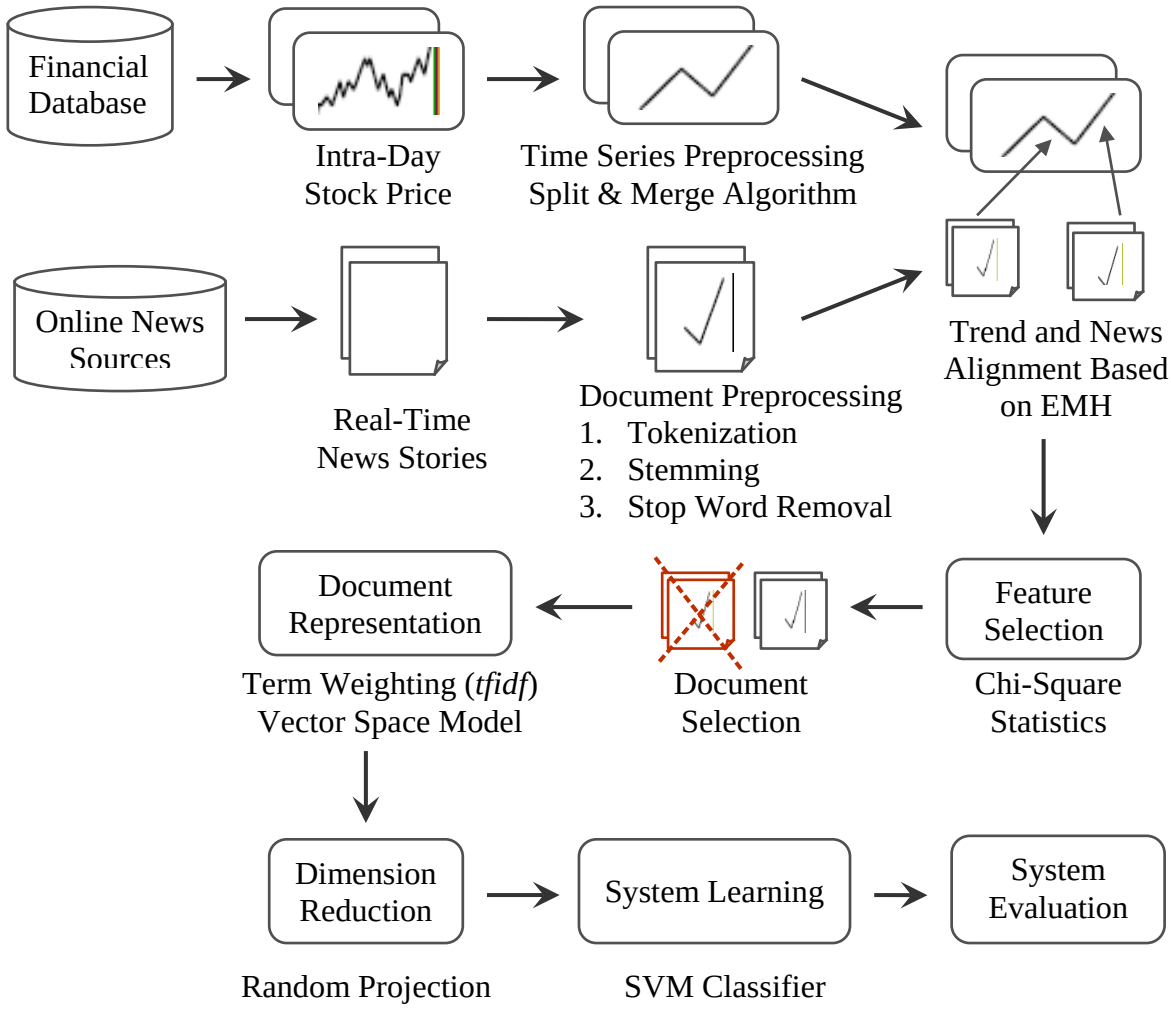


Figure 5.2: The Overall Research Process

In order to accomplish the overall research process, an extensive programming is required. We have used R Open Source Programming Language and Python Programming Language to implement the overall research process. R is a language and environment for statistical computing and graphics which provides a wide variety of statistical (linear & nonlinear modeling, time-series analysis, classification, clustering …) and graphical techniques, and is highly extensible. It is an environment within which statistical techniques are implemented and can be extended (easily) via *packages*. (R-Project for Statistical Computing, 2003) We have also used Python which is a dynamic object-oriented programming language that can be used for many kinds of software development. (Python Programming Language, 1990) The algorithms related to each step is programmed and coded in one of the above programming environment. The different programming are then combined together to make the overall prediction package. In order to summarize the research process, we have provided Figure 5.3 which shows the end result of the prediction package.

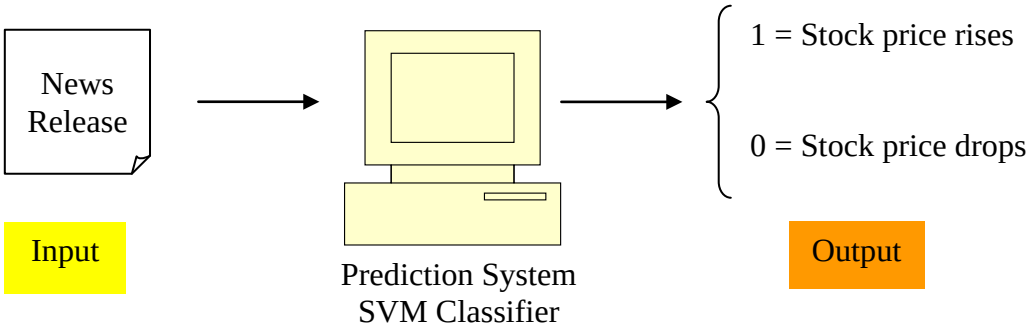


Figure 5.3: The Prediction Model

Classifiers for various categories (Up and Down) are induced in the system learning phase of the research process and the prediction model is made using SVM classifier. The aim of the prediction model is to determine the impact of an upcoming news release on the stock price trend. Hence the news which is released from a news agent is fed to the prediction model as an input. The output of the system could be either 0 or 1. If it is 0, it means that the news has the negative impact on the stock price and the stock price will drop and if it is 1, the news will cause an increase in the stock price. In either case the investor can act upon it and make a right decision.

5.2.1 Data Collection

For conducting our study, two types of data are needed: the intra-day stock price and time-stamped news articles. In the following sections, we describe the way these data are collected.

Intra-day Stock Price Collection

As the name implies, intra-day stock prices includes all prices of a single stock within a day. For predicting the trend of stock prices, we need the records of intraday prices of certain stocks over a specific time period. These data are the internal secondary data stored in the databases of financial institutes or organizations. We made a broad investigation to find the best and the most reliable sources of data. By referring to some stock brokers and explaining the type of data required, they introduce Tehran Stock Exchange Technology Management Company (TSESC) as the only source which has the archive of all stock prices. We chose the 20 most active stocks during years 1383 and 1384 with the help of a consultant in Bank Sepah Stock Broker. We prepared a request and identified the exact data specifications. Tehran Stock Exchange Technology Management Company (TSESC) provided 20 text files each related to a specific stock. These files are consists of different columns including the name of the stock, the date, the time (hour and minute), the intraday price and the volume traded starting from the first working day in year 1383 to last working day in year 1384.

Time-Stamped News Articles

The other data required for this study is the collection of time-stamped news stories about the companies whose stock price movement is going to be predicted. News articles are the secondary data which can be either internal stored in some organizations' databases or can be external and be available in the internet. As the purpose of this study is to investigate the financial behavior of investors after news releases, we focus on extracting time-stamped news articles from online sources. The overall procedure can be divided into 3 parts: finding the news sources, gathering news links from selected websites, and saving news from the links.

Finding News Sources

A comprehensive browsing on the internet is done to find out if there exist any online sources which provide archived time-stamped news articles. After finding these types of websites exist, we started to find almost all the online news sources and websites which provide an archive of financial and economical news about Iranian companies and organizations. We realized that these news sources fall under one of the following categories including news agencies, news sites, financial and stock related sites, financial newspapers, non-financial newspapers, economic sites, and stock web logs. Then the most important web sites for each category are found and saved by their names and links in an Excel sheet. Each website is fully examined for some criteria and those that do not provide time-stamped news are omitted from the list. Among different online news sources, only 15 were selected as the source of news gathering, twelve of which belong to the category of news agencies, 3 of them belong to news sites, and one belongs to financial newspapers. Refer to Appendix 1 to find out more about news sources and the number of news gathered from each of them.

Gathering News Links from the Selected Websites

Unfortunately there is no enough real-time news about most of the companies whose stock prices are gathered. Among the 20 companies, Iran-Khodro found to be the only company whose news articles are almost enough for the purpose of this study. An Excel sheet is prepared for news related to Iran-Khodro during the years 1383 and 1384. This Excel sheet includes the news release time, the news links, and the news sources. Refer to Table 5.1 to see an example of news links and their release time. The news release time is inserted in the first column manually in a 10 digit format. (The year, the month, the day, the hour and the minute all in a 2-digit format from left to right, yymmddhhmm). The news links are saved in the second column. The links to be saved is better to be the printable page links as these pages have the characteristics of containing only the date and time, the title, and the body of the text and no excessive material is included in printable pages. The source of news is inserted in the third column. From 15 selected online news providers, 1523 time-stamped financial, economical, and political

news related to Iran-Khodro were gathered. The details of surfing and gathering news from these web sites differ from one another. Most of these online news sources have provided a simple or advanced search engine which can be used for finding the relevant news. The Iran-Khodro (ایران خودرو) term is used as the main keyword. The other keywords used include Paykan (پیکان), Samand (سمند), and Peugeot (پژو). Some of the search engines only accepted one single word, and we had to search for the term Khodro (خودرو) and then manually choose the ones related to Iran-Khodro Company. The advanced search engines provided the facility to search in the specified time period (1383 and 1384), to search in different news category (political, financial, economic, industrial, and commercial), and to search in different parts of the news (title, subtitle, and body of the text). News with the specified criteria had to be searched manually for those websites who only provides simple search engine. We only took into account the news exactly related to Iran-Khodro activity, and the ones related to automobile industry and other car manufacturer are not considered in our study.

Table 5.1: Examples of News Links and Their Release Time

Date & Time	Link	Source
8412231204	http://www.ilna.ir/print.asp?code=291118	ILNA
8412201542	http://www.ilna.ir/print.asp?code=290115	ILNA
8412161506	http://www.ilna.ir/print.asp?code=288836	ILNA
8412141041	http://www.ilna.ir/print.asp?code=287696	ILNA

Extracting News from the Links

After saving the news links, news should be saved in text format. 1523 text files are saved each containing the title and the news text. The news files are named by their two letter abbreviation indicating the source of the news, and their 10-digit date and time number. For example, IL8412231204 indicates that the news is gathered from ILNA news source and is released in 23rd of the 12th month in 1384 at 12:04. At the end, we would have 1523 text files of Iran-Khodro news and their corresponding time releases. These names of these news files are brought in Appendix 2.

5.2.2 Document Preprocessing

The document preprocessing consists of 3 steps including tokenization, stemming and stop word removal. Refer to Chapter 4 for the complete explanation. All the processes in document preprocessing are done with Python Programming Language. Three different programs are written to perform tokenization, stemming and stop word removal over the news files. In tokenization process, all the punctuations, numbers, and extra marks are first removed from the news texts. There is a list of Persian spelling which automatically checks the spelling of the news texts. Then the body of each text is divided to some number of words. The program will recognize each word at the place where a space is entered in the text. After all news texts are tokenized, stemming is performed on each word in order to transform it to its rooting form which is done based on a list of Persian suffix and prefix (a simple form of Porter's Algorithm). Then stop words (words with high frequency such as conjunctions) are removed based on the stop word list (domain dependent) which has been prepared by reviewing all the words extracted in stemming process. (Refer to Appendix 3 for the list of stop words) The rare words (those repeated less than 5 times) will be removed in the feature selection process.

5.2.3 Time Series Preprocessing

As with most data mining problems, data representation is one of the major elements to reach an efficient and effective solution. Since all stock time series contains a high level of noise, a high level of time series segmentation is necessary for recognizing the significant movements or detecting any abnormal behaviors, so as to study its underlying structure. Piecewise linear segmentation, or sometimes called piecewise linear approximation, is one of the most widely used technique for time series segmentation, especially for financial time series. The details of techniques are fully explained in Chapter 3. We have used t-test based split and merge algorithm for segmenting the time series proposed by Fung et al. (2005) Fung has used the confidence level of 0.95 percent but due to the nature of our time series we have used 0.999 as the confidence level to avoid over segmentation. We used R Open-Source Language Programming for implementing the split and merge algorithm and Minitab to draw the original and split

time series. The algorithm consists of two parts: the splitting phase and the merging phase. The splitting phase aims at discovering trends on the time series, while the merging phase aims at avoiding over-segmentation.

Splitting phase: Initially, the whole time series is regarded as a single large segment. It is represented by a straight line joining the first and the last data points of the time series. To decide whether this straight line (segment) can represent the general trend of time series, a one tail t-test is formulated:

$$H_0: \epsilon = 0 \qquad H_1: \epsilon > 0$$

ϵ is the expected mean square error of the straight line with respect to the actual fluctuation on the time series. The square sum of the distance between all data points within the segment and the regression line is calculated. k is the total number of data points within the segment, $\hat{p}_i$ is the projected price of p_i at time t_i.

$$\varepsilon_i = (p_i - \hat{p}_i)^2 \qquad \varepsilon = \frac{1}{k}\sum_{i=0}^{k}(p_i - \hat{p}_i)^2$$

The t-test is performed on the mean square error calculated. If the null hypothesis is accepted (p-value > α = 0.001), then the mean square error between the actual data points and the projected data points should be very small and the straight line, which is formulated by joining the first and the last data points of the segment, should be well enough to represent the trend of the data points in the corresponding segment. In contrast, if the null hypothesis is rejected (p-value < α = 0.001) and the alternative hypothesis is accepted, then that straight line is not well enough to represent the trend of the data points in the corresponding segment and should be split at the point where the error norm is maximum, i.e. $\max_i \{(p_i - \hat{p}_i)^2\}$, and the whole process will be executed recursively on each segment until condition (p-value > α = 0.001) holds for all the segments.

Merging Phase: Over-segmentation will frequently occur after the splitting phase. Over-segmentation refers to the situation where there exist two adjacent segments such that their slopes are similar, and they should be merged to form a single large segment.

Merging aims at combining all of the adjacent segments, provided that the mean square error for each adjacent segment would be accepted by the t-test. If the null hypothesis over two adjacent segments is accepted (p-value > $\alpha = 0.001$), then these two segments are regarded as a merging pair. The hypothesis for the t-test is the same as the ones in splitting phase. For all the adjacent segments the mean square error would be calculated. The t-test over all the error norms would be performed. Those whose null hypothesis is accepted are regarded as merging pairs. The program start merging from the segments whose mean square error is the minimum. When two adjacent segments are merged, the new segment should be checked with its previous and next segment and the mean square error for them would be calculated and t-test is performed over them. If any null hypothesis is accepted the corresponding error norm would be added to the list and again among the mean square errors the minimum would be chosen to merge the corresponding segments. The whole process is executed continuously until the t-test over all of the segments on the time series is rejected and there is no merging pair left.

5.2.4 Trend and News Alignment

After document and time-series preprocessing, news articles should be aligned to trends. By alignment, we mean that the contents of the news articles would support and account for the happening of the trend. For aligning news stories to the stock time series, there could be three different formulations under different assumption: (Fung et al., 2005) let us take Figure 5.4 to illustrate these ideas.

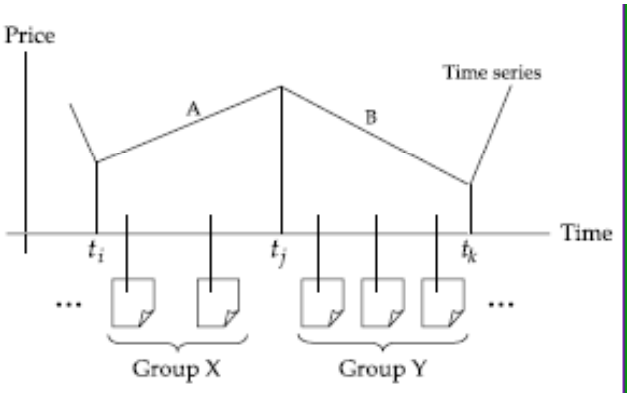


Figure 5.4: News Alignment Formulation

Source: Fung et al., 2005

Observable Time Lag: In this formulation, there is a time lag between the time when news story is broadcasted and the movement of stock prices. It assumes that the stock market needs a long time for absorbing the new information. Refer to Figure 5.3. In this formulation, the Group X (news stories), is responsible for triggering Trend B. Some reported works have used this representation including Lavrenko et al. (2000), Permunetilleke and Wong (2002), and Thomas and Sycara (2000).

Efficient Market: In this formulation, the stock price moves as soon as after the new story is released. No time lag is observed. This formulation assumes that the market is efficient and no arbitrary opportunity normally exists. Refer to Figure 5.3. Under this formulation, Group X is responsible for triggering Trend A, while Group Y is responsible for triggering Trend B.

Reporting: In this formulation, new stories are released only after the stock price has moved. This formulation assumes that the stock price movements are neither affected nor determined by any new information. The information (news stories) is only useful for reporting the situation but not predicting the future. Under this formulation, in Figure 5.3, Group Y is responsible for accounting why Trend A would happen.

Different scholars may be in favor of one of the above formulations. But there is no clear distinction as which formulation performs better and they have not reached to a common consensus. The choice of formulation may be dependent on the nature of the market under study. In our methodology the alignment is based on the Efficient Market Hypothesis which is completely discussed in Chapter 2. The alignment process would be accomplished by a program written in R Language Programming. The program will check each news story (d_i) release time (t_{rel}) and compare it with the beginning (t_{begin}) and ending time (t_{end}) of each segment (S_k). The document will be assigned to a segment whose time release is between the beginning and ending time of a particular segment. D denotes all of the news stories archived and D_{Sk} denote the documents that are aligned to Segment S_k. We can summarize the alignment query as follows:

$$di \in \{D_{Sk} \mid t_{rel}(di) \geqq t_{begin}(S_k) \text{ and } t_{rel}(di) < t_{end}(S_k)\}$$

5.2.5 Feature and Useful Document Selection

After aligning news articles to trends, we need to select useful documents. In reality, many news stories are valueless in prediction; it means that they do not contribute to the prediction of stock prices. Hence we have to filter out news articles that do not support the trends and keep the useful ones. The usefulness of a document can be determined by the features it contains. If any of the features in a document (news story) is said to be significant in the stock trend movement, then the documents containing those features are the ones contributing to the stock price prediction. The methods of feature selection in text categorization and their evaluations are completely reviewed in Chapter 4. Our selection of features and news stories is based on a χ^2 (Chi-Square) estimation on the keywords distribution over the entire document collection. Feature and document selection algorithm is going to be implemented in Python Language Programming. Before explaining the algorithm we first briefly introduce the concept of chi-square statistic.

Chi Square Statistic (CHI)

In text analysis, the statistically based measures that have been used have usually been based on test statistics which are useful because, given certain assumptions, they have a known distribution. This distribution is most commonly either the normal or chi-square distribution. These measures are very useful and can be used to accurately assess significance in a number of different settings. (Dunning, 1993)

The chi square statistic measures the lack of independence between a term (t) and a category (c) and can be compared to the chi-square distribution with one degree of freedom to judge extremeness. (Yang and Pedersen, 1997) For each term, a 2x2 contingency table is constructed to determine its corresponding chi-square value. With the model that tokens are emitted by random process two hypotheses should be assumed. First the random processes generating tokens are stationary, meaning that they do not vary over time, and second the random processes for any pair of tokens are independent. If the process producing token (t) is stationary, then for an arbitrary time period t_0 the

probability of seeking the token is the same as the probability of seeing the token at other times. The assumption that two features t_i and t_j have independence distributions implies that $P(t_i) = P(t_i \mid t_j)$. (Swan and Allan, 1999)

Using the two-way contingency table of a term f_j and a category S_k, we can calculate the value of chi-square for each term. (Table 5.2) A is the number of documents that contains feature f_j and is in segment S_k, B is the number of documents that contains feature f_j but is not in segment S_k, C is the number of documents that does not contain feature f_j but is in segment S_k, and D is the number of documents that does not contain feature f_j and is not in segment Sk. N is the total number of documents.

Table 5.2: A 2x2 Contingency Table; Feature fj Distribution in Document Collection
Source: Fung et al., 2005

	# Documents have f_j	# Documents do not have f_j
Segment = S_k	**A**	**C**
Segment ≠ S_k	**B**	**D**

The term-goodness measure is defined as the following formula. The chi-square statistic has a natural value of zero if the feature and the category are independent. (Yang and Pedersen, 1997) The larger a χ^2 value, the stronger the evidence that term and category are dependent on each other and the occurrence of feature in that category is significant.

$$\chi^2(t,c) = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)}.$$

After aligning the documents to time series segments, we computed for each category the chi-square statistic between each unique term in a training corpus and that category based on the above formula. Note that for $\chi^2 = 7.879$, there is only a probability of 0.005 that a wrong decision would be made such that a feature from a stationary process would be identified as not stationary, means that a random feature is wrongly identified a significant feature. Hence if the chi-square value is above 7.879 it is concluded that the term's appearance on that segment is significant and this is the measure used by Fung. On the other hand they calculated the χ^2 value for those features

that appear in more than one-tenth of the documents means that A/ (A+C) > 0.1. We changed these measures due to the nature of the features. If we choose $\chi^2 = 7.879$ almost all of features are considered as significance, and few of them would be removed from the feature set. Hence we changed the chi-square value to a higher level so that the most significant ones can be chosen. We set this value equal to 10 ($\alpha = 0.001$). Besides, we have only calculated the chi-square values for features that appear in more than two-tenth of the documents in the corresponding segment means that A/ (A+C) > 0.2. If any cell in the contingency table is lightly populated, which is the case for low frequency terms, the chi square statistic is known not to be reliable and selecting terms by using χ^2 when the cell value is less than 5 have been criticized for being unreliable. (Dunning, 1993) Hence those features whose contingency table values are equal or more than 5 are going to be considered. The features, whose chi-square values are above 10, are appended into a feature list related to their segment. Based on the features selected for each segment, the useful documents are going to be selected. Any of the documents in a segment which contains any of the features related to that segment is going to be chosen as a useful document. Those documents that do not contain any of the selected features are going to be discarded. The documents selected are going to be classified into two main groups, those belong to the rising segments D_R, and those belong to the dropping segments D_D.

5.2.6 Document Representation

As mentioned in Chapter 4, the simplest and almost universally used approach in document representation is the bag-of-words representation. We have used *tfidf* as the term weighting method in our vector space modeling. Each document in both D_R and D_D is going to be represented by a vector of numeric values, each value corresponding to the term's importance in that document which is calculated by the *tfidf* formula and the features that are not in that document get the value of 0 in the representation. Hence each document has n-dimension (R^n), where n is the total number of features in D (selected news). Weights obtained by *tfidf* equation are then normalized to unit length by cosine normalization to account for the differences in the length of each document. We have used Python Programming to implement the representation process.

5.2.7 Dimension Reduction

In Chapter 4, we made a comparison between the two types of dimension reduction techniques, namely feature selection, and feature transformation (extraction) methods. After reducing the feature set size using the chi-square feature selection and selecting the useful documents, the dimensionality of the represented documents is still too high to be accepted by the SVM classifier. In order to reduce the size and dimensionality of the matrix constructed in the document representation process, we have to apply one of the feature transformation techniques.

Feature transformation methods perform a transformation of the vector space representation of the document collection into a lower dimensional subspace, where the new dimensions can be viewed as linear or non-linear combinations of the original dimensions. (Tang et al., 2005) An ideal dimensionality technique has the capability of efficiently reducing the data into a lower-dimensional model, while preserving the properties of the original data. One common way to reduce the dimensionality of data is to project the data onto a lower-dimensional subspace. (Lin and Gunopulos, 2003) Among the popular techniques mentioned in Chapter 4, we chose random projection (RP) to reduce the dimensionality of our represented documents. In the following section, we briefly explain the idea behind the random projection mapping. The random projection is programmed and implemented in Python Programming Language.

Random Projection

Random projection is a powerful technique for dimensionality reduction. The method of random projection (RP) was developed to provide a low (computational) cost alternative to LSI for dimension reduction. Naturally, researchers in the text mining and information retrieval communities have become strongly interested in RP as it has been proven to be a reasonably good alternative to LSI in preserving the mutual distances among documents. (Bingham and Mannila, 2001) Random projection can be applied on various types of data such as text, image, audio, etc. It is based on a simple idea and is efficient to compute (Lin and Gunopulos, 2003).

The key idea of random mapping arises from the Johnson-Lindenstrauss lemma (Johnson and Lindenstrauss, 1984) which states that if points in a vector space are projected onto a randomly selected subspace of suitably high dimension, then the distances between the points are approximately preserved. The method of random projection is a simple yet powerful dimension reduction technique that uses random projection matrices to project the data into lower dimensional spaces. It has been shown empirically that results with the random projection method are comparable with results obtained with PCA, and take a fraction of the time PCA requires. (Fodor, 2002)

The idea behind random projection is simple. It states that given the original matrix $X \epsilon R^m$, the dimensionality of the data can be reduced by projecting it through the origin onto a lower-dimensional subspace, matrix $A \epsilon R^k$ with k<<m. This is done via a randomly generated projection matrix R. (Fodor, 2002)

$$A_{[k\times n]} = R_{[k\times m]} \cdot X_{[m\times n]}$$

Several algorithms have been proposed to generate the matrix R. The elements of matrix *R* can be constructed in many different ways. One can refer to Deegalla and Bostrom (2006) to read more on generating the random matrices.

We have used a very simple form of random projection. We decided to reduce the dimension of the represented documents from m = 4839 to k = 200. For doing so we have multiplied the original matrix X of n = 447 and m = 4839 by a random matrix of m = 4839 and k = 200. The new matrix has 447 rows (the number of documents) and 200 columns which are the number of features. Our random matrix has the elements of 0 and 1. The number of element 1 is set to 5 in the programming, means that in each column of the random matrix 5 cells are chosen randomly and are assigned number 1 and the rest of the cells are of 0 value. Each row in the original matrix is multiplied by 200 different columns each including 5 elements of value 1 selected randomly. The overall process of constructing the new matrix is the product of 404 rows each multiplied by 200 columns. After constructing the new matrix, the elements of the matrix would be normalized to unit length and would be ready to be given as input to the text classifier.

5.2.8 Classifier Learning

The relationship between the contents of news stories and trends on the stock prices are learned through support vector machine classifier (SVM). The major learning and prediction process is implemented in R Language Programming using Package e1071. The new document representation (tfidf) after dimension reduction has 447 documents each represented by 200 features. Except the 200 columns, the first column is the labeling of the documents. Each document is labeled 1 if it is in documents responsible for rise event and is labeled 0 if it is in documents responsible for drop event. We need no class balancing as the number of the drop events is almost two times the number of rise events. The machine learning approach relies on the existence of an initial corpus of documents previously categorized as up or down. For evaluation purposes in the first stage of classifier construction, it is needed that the initial corpus is divided into two sets, namely the training and the test set. The training set is the set of example documents observing the characteristics of which the classifiers for various categories are induced. The test set will be used for the purpose of testing the effectiveness of the induced classifiers. Each document in test set will be fed to classifiers, and the classifier decision would be compared to the actual or expert decision. We have randomly chose 70 percent of our documents as the training set and 30 percent as the test set. Using the SVM classifier, we have to identify the type of kernel that is used in classification process. Among the existing kernels (Chapter 4), we chose RBF kernel which is proved to be the best for text classification problems. Using the RBF kernel, we have to set the two parameters associated with this kernel namely the cost (c) and the gamma (γ). The details of setting these two parameters are fully discussed in Chapter 6.

5.2.9 System Evaluation

The methods for evaluating the performance of text classifiers have been discussed in Chapter 4. Our classification effectiveness has been evaluated using classifier accuracy, precision-recall curve, precision-recall F-measure, and ROC curve. All of these evaluation measures are induced from the confusion matrix given by the SVM classifier. The ROCR Package in R Language has been used to draw the curves.

Chapter 6

Results and Analysis

6. Results and Analysis

The results related to different steps of the research process are provided in this chapter. All of these processes are implemented and conducted either in R or Python Programming Language. The results obtained from the text classifier (SVM) are then analyzed and the prediction model and the overall system performance would be evaluated using different evaluation measures.

6.1 Time Series Segmentation Results and Evaluation

Among the 20 text files provided by Tehran Stock Exchange Service Company (TSESC), Iran-Khodro intraday price and their corresponding date and time during years 1383 and 1384 are chosen to be given to the split and merge algorithm. Before implementing the segmentation algorithm in R Programming Language, Iran-Khodro text file should be read by the program. The program reads 46232 intraday prices (data points) for this company during years 1383 and 1384 which is equal to 46231 segments in Iran-Khodro time series plot. The file is given to the split algorithm which reduces the number of segments from 46231 to 4777. Then these 4777 segments are given to merge algorithm and 1811 segments are produced. The segmentation algorithm reduces the total 46231 segments to 1811 segments.

In order to evaluate and illustrate the segmentation process, we exported the original, split, and merged data to Minitab Statistics Software to draw the time series graphs. Figure 6.1 illustrates the original time series including 46232 data points while Figure 6.2 is the segmented time series having only 1812 data points. The segmentation algorithm has reduced the number of data points from 46232 points to 1812 points.

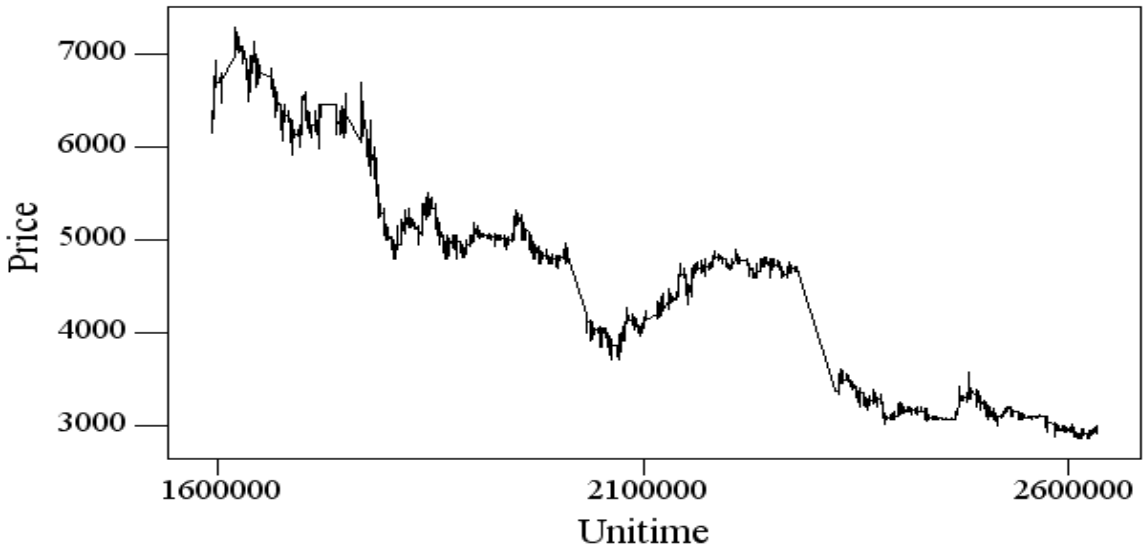


Figure 6.1: Iran-Khodro Original Time Series for Years 1383 and 1384

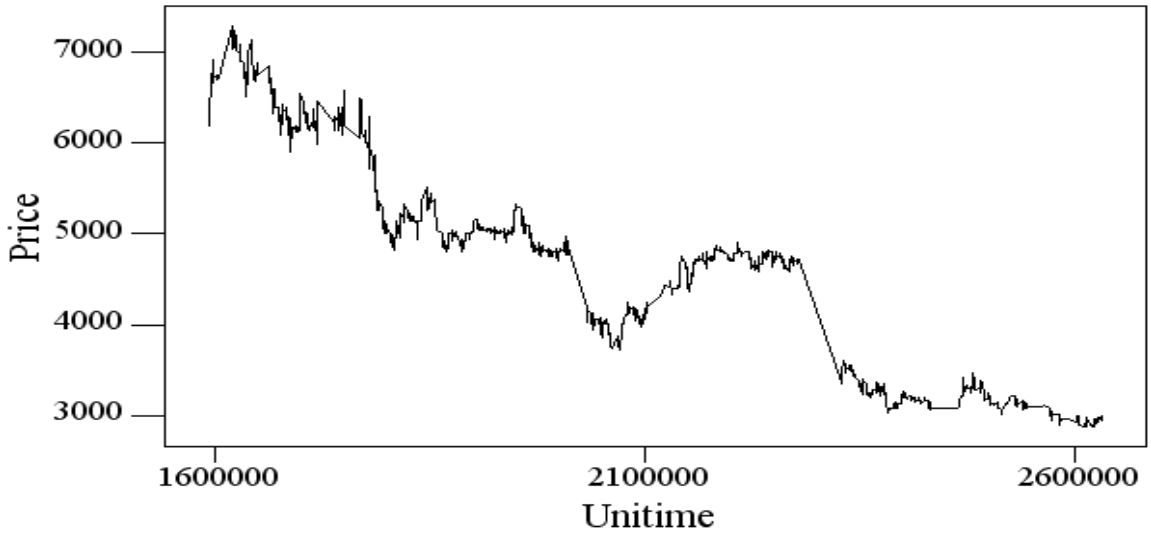


Figure 6.2: Iran-Khodro Segmented Time Series for Years 1383 and 1384

Because of the density and compactness of data, one may still not recognize the difference between Figure 6.1 and 6.2. As a result, we took a small sample period of time series to better illustrate the effect of split and merge algorithm on Iran-Khodro time series. These illustrations are provided in Figure 6.3 and 6.4.

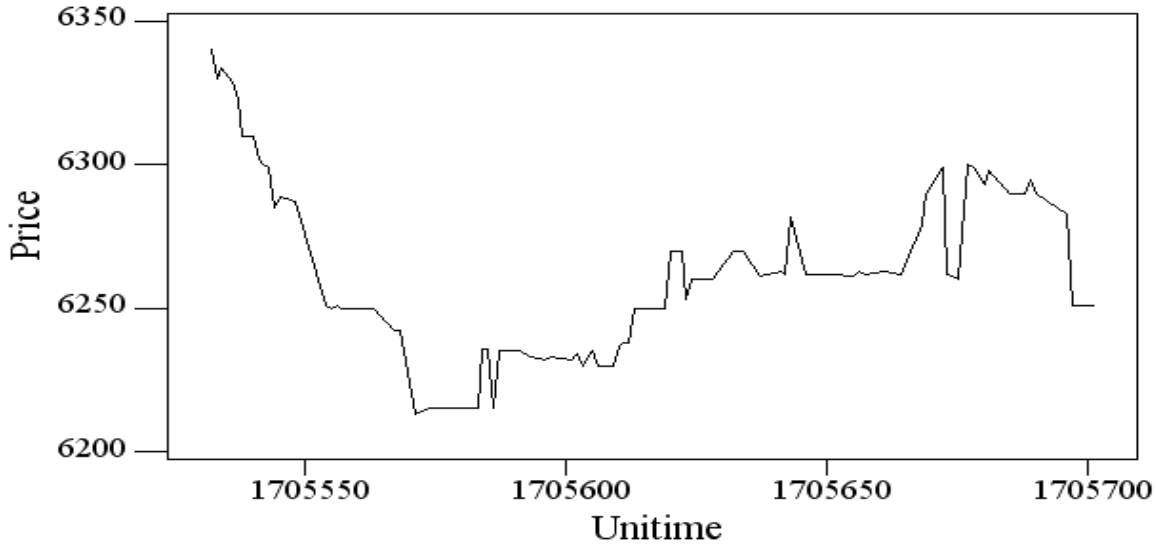


Figure 6.3: Iran-Khodro Original Time Series; Small Sample Period

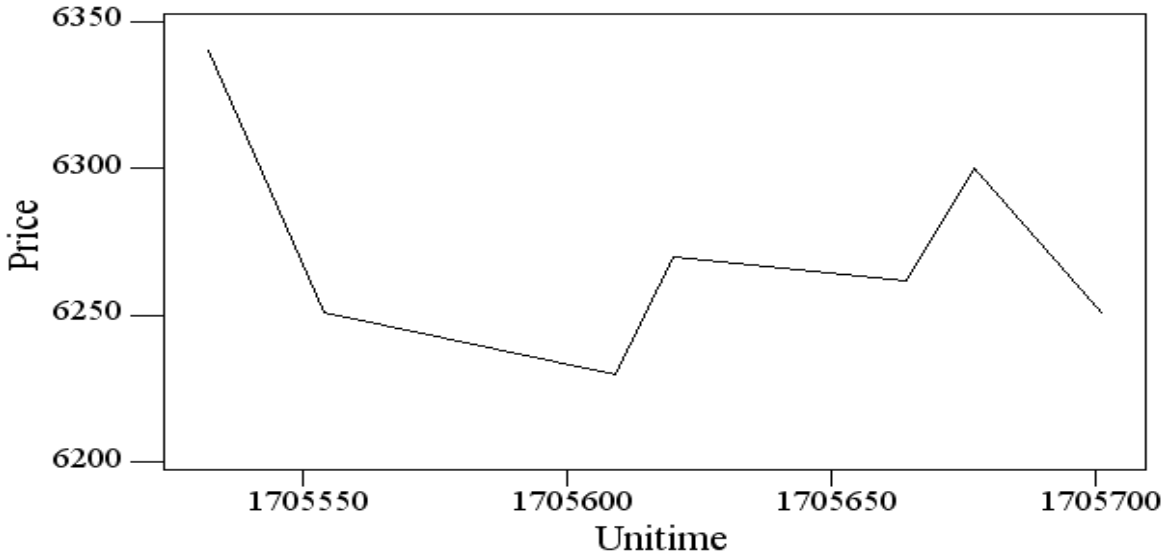


Figure 6.4: Iran-Khodro Segmented Time Series; Small Sample Period

6.2 News and Trend Alignment Results

1523 financial and political news (Refer to Appendix 2) are gathered about Iran-Khodro Company during years 1383 and 1384. Before the alignment, all of the news is first preprocessed. Out of these 1523 news only 1516 news are aligned back to segments. 7 pieces of news are not aligned back to trends as their release time was either less than the beginning time of the first segment or more than the end time of the last segment (Those whose release time is less than 8301080908 and bigger than 8412281227). Green cells in Appendix 2 specify the news whose release time is out of the scope. 1516 news is then aligned back to 1811 segments resulted in time series preprocessing. Out of 1811 segments, only 429 segments received news. Refer to Appendix 3 for the news and trend alignment results. 717 pieces of news belong to rise-trends and 799 belong to drop-trends which indicate the alignment process is almost balanced between the two trends.

6.3 Document Selection & Representation Results

After news is aligned back to trends, useful documents should be selected. The chi-square feature selection program which is coded in Python Programming Language is conducted on the 1516 aligned documents which contain the total number of 8980 features. 294 features are selected as significant features in feature selection process. Refer to Table 6.1 for the selected features in rise and drop trends. Any of the documents that contain any of the significant features in any segments would be chosen as useful documents. The total number of selected documents is 447 pieces of news and the 1069 news are discarded. The 447 selected documents are represented using tfidf weighting scheme. Python Programming is used to implement the representation process. The result of representation is an Excel sheet with 447 rows each corresponding to a document and 4839 columns indicating the total number of features in selected documents. There exists another column which identifies the category to which the news belongs. Hence, each document is going to be represented by a vector of numeric values, each value corresponding to the term's importance in that document which is calculated by the *tfidf* formula. Features that are not in that document get the value of 0 in the representation. Weights are normalized by cosine normalization and get the value between 0 and 1.

Table 6.1: Selected Features for Rise and Drop Segments Using Chi-Square Metric

Features for Drop Trends				Features for Rise Trends		
هزار	ساز	تجهیزات	محصولات	جمع	مدیر	عمده	حقوق
تنوع	محصولات	توضیح	جهان	دارندگان	ساز	مدیره	محصولات
جدید	فروش	درصد	کمك	تهران	تعیین	ترکیب	سهامداران
موتور	تغییرات	زیاد	دارنده	صندوق	دستگاه	اعضا	عاد
وارد	دستگاه	هنگام	قدیم	پیکان	جهان	عضو	فوق
ترمز	طراح	ساز	غیر	مالك	مجمع	هزار	فروش
حفظ	برند	هزار	ازا	شامل	مجموع	تیراژ	میلادی
کاهه	طرح	خریدار	دستگاه	اولو	مدیره	مدیریت	برخوردار
تکمیل	مدیر	سید	بودجه	معادل	مثبت	سریع	نظارت
یابد	مراسم	ابراز	لیزینگ	توسط	برطرف	شورا	گام
درصد	زان	گفتوگو	ماهه	کارمزد	نکته	حما	معتبر
مبزی	گیر	اقساط	طرح	استقبال	کالا	هزار	ساز
تعرفه	جدید	انتخاب	کیف	نهاد	حجم	مقایسه	صندوق
نیرو	هزار	جدید	اختیار	ذیربط	موتور	طراح	برند
ساز	دهند	فروش	محصولات	منظور	مجموعه	دستگاه	صادرات
طرح	محصولات	آزاد	یافته	تشریح	اروپا	آریان	برخوردار
نصب	احتمال	شروع	وانت	مراجعه	طرح	رونما	محصولات
انتظامی	پلیس	پشتیبان	ممکن	لا	نقل	جدید	استاندارد
ترمز	خبرنگاران	درصد	حسابرس	دو	ماه	تیپ	عنوان
روزانه	فروش	طرح	قانون	عملکرد	دستگاه	طراح	درون
یادآور	دهناد	مربوط	مطلوب	دستاورد	ساز	موتور	هزار
طراح	تخصص	جدید	عادی	مربوط	طرح	خدمات	داخل
موتور	نخستین	حال	مجمع	محصولات	منطقی	مدیر	محصولات
گازسوز	بنزین	مدیر	درصد	مرکز	همایش	صادرات	دانش
نشست	درصد	کاهش	افزا	طراح	صنایع	فروش	قرار
نما	زدیدکنندگان	کیف	هزار	ساز	چهارمین	کیف	درصد
دهند	استقبال	انجام	جدید	صورت	علت	طراح	مدت
کار	سالانه	دستگاه	قرار	بین	شبکه	ارتقا	صورت
مدیر	صادرات	اهداف	فروش	محصولات	برقرار	جلسه	نظارت
کیف	تبدیل	صدور	برخوردار	محور	وضع	لحاظ	ضمن
صورت	دستگاه	اقتصاد	افزا	درصد	عضو	برطرف	محصولات
حال	دو	نقل	جدید	استاندارد	مرتبط	فروش	کمیسیون
هزار	ید	اعلام	طراح	روابط	ان	زاریاب	محصولات
عنوان	صنایع	فروش	عموم	قرار	فروش	دو	امور
مرکز	اشاره	قیمت	خارج	خودروساز			
ساز	طرح	خدمات	منطق	داخل			
جهان	ماه	محصولات	آینده	انجام			
ظاهر	دوم	سر	کیف	تعویض			
مالی	زالی	کلاس					

Due to the huge volume of *tfidf* document representation, it was not possible to show the entire result. Hence, we have illustrated a sample of our document representation result in Table 6.2. This is a minimized excel sheet with 447 rows and 4840 columns. This representation is actually our training data to be given to SVM classifier. The first column (purple color) as mentioned before identifies the category of the selected documents either by 0 as drop trend or 1 as rise trend. In front of each document, there is a representation of 4839 numeric values which are either 0 for features that are not included in the document, or a value between 0 and 1 for the features that are included in the document. The green area shows an example of the *tfidf* representation and weighting scheme.

Table 6.2: An Illustration of tfidf Document Representation

		Rise(1) Drop(0)	**4839 = Total Number of Features in Selected Documents**											
		1	F1	F2	F3	F4	.	.	.	.	.	.		F4839
447 Selected Documents	1	1	0	0	0.13	0	0	0	0	0	0	0.76	0	0
	2	0	0.66	0	0	0	0	0	0	0	0.14	0	0	0.10
	3	0	0	0	0	0.86	0	0	0	0	0	0.32	0	0
	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	446	1	.	.	.	.	.	.	.	.	.	.	.	.
	447	1	0.22	0	0	0	0	0	0	0	0.18	0	0	0.78

6.4 Random Projection Result

We gave the tfidf representation as input to the SVM classifier. The number of features which is equal to 4839 is still too large and SVM cannot handle this amount of features. Hence we reduced the dimension of the features using random projection technique. Here we have not omitted the features, but instead we have made a combination of them which is discussed in Chapter 5. We reduced the number of columns in tfidf representation to 200 combined features which is implemented in Python.

6.5 Classifier Learning and SVM Results

The association between different news and different trend types are generated based on Support Vector Machines. The tfidf representation is given to SVM classifier as an input. R Language Programming (Package e1071) is used to implement the classification process. We have chosen radial basis function kernel (rbf) as it is yields the best performance in text categorization problems. For all of the news aligned to a stock, 2/3 of them are randomly grouped into a train set (298 news) and the remaining 1/3 are grouped into a test set (149 news). The prediction model is built according to the train set, which is then used to evaluate the prediction model by the test set.

Certain number of parameters should be set for using SVM classifier in R Language Programming, the cost (c), and the gamma (γ). Cost (misclassification error) and gamma should be set in such a way that minimize the error of the classifier and improve its accuracy. In order to determine the best cost and gamma for the classifier, a five-fold cross validation experiment is conducted on a range of numbers considered for each parameter. The program will classify the data with the different combinations of cost and gamma 5 different times and choose a pair which results in lower error rate on the train set.

The result of such combinations is illustrated in Figure 6.5 which is called the performance of SVM. In this figure, X axis represent gamma (RBF parameter) and Y axis represent the cost (SVM parameter). The model is built on the train set for each of these data points in the figure (different combinations of cost and gamma), and the error rate associated to each model is calculated and shown with different colors. As it is shown in the picture, the white area has the lowest error rate which indicates the model best performance. The selected cost and gamma lies in the white area with the minimum error rate of 23%. The model best performance is when cost is equal to 3.583 and gamma is equal to 1.728 which is shown on Figure 6.5. We give the training data to the SVM classifier with the mentioned cost and gamma, and the prediction model would be built accordingly. After model is built, the test set (149 news) is given to the model to evaluate the performance of the classifier. The results are given in the following section.

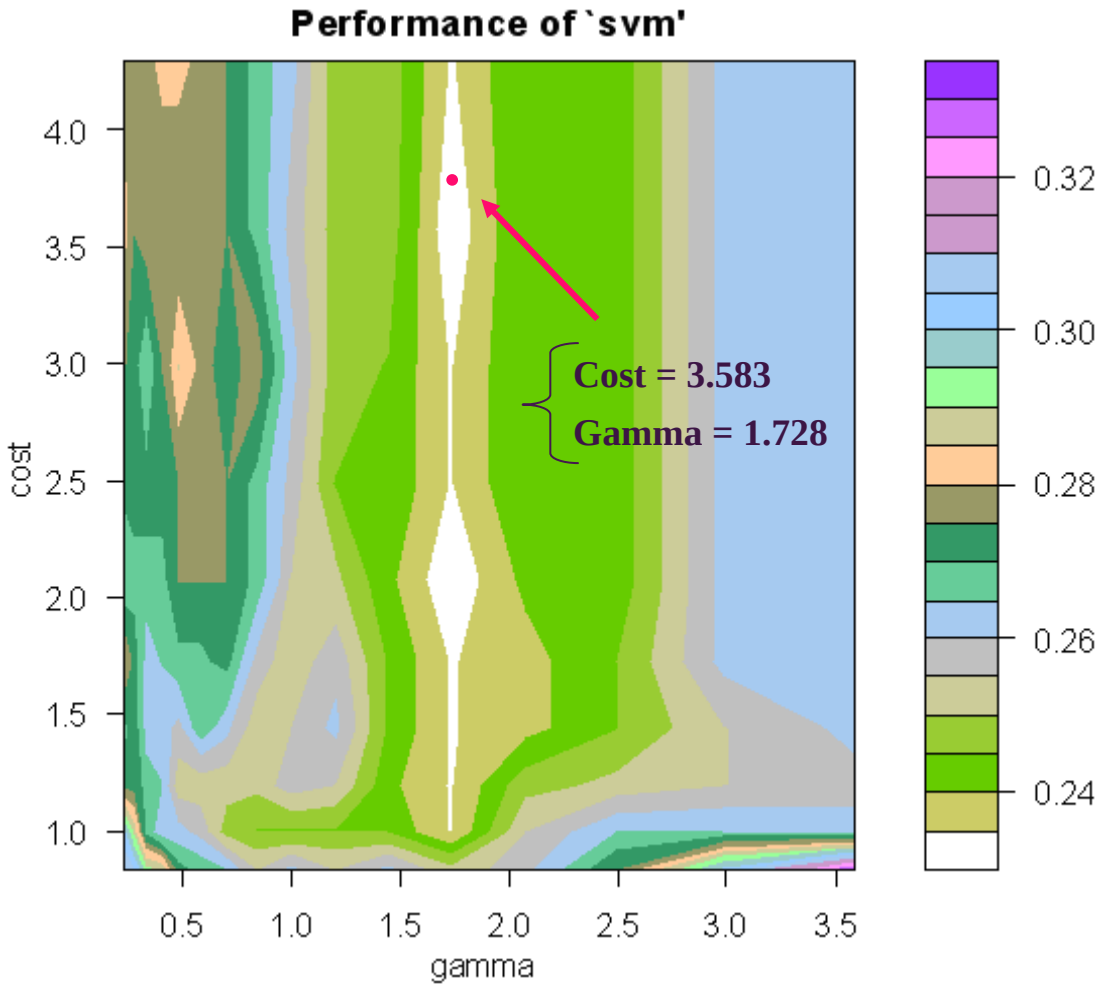


Figure 6.5: SVM Parameter Tuning

Result of Prediction Model

The decision made by the classifier can be represented in a structure known as a confusion matrix or contingency table which is explained completely in Chapter 4. The confusion matrix for model prediction is provided in Table 6.3 which is then used for the evaluation of classifier performance.

Table 6.3: Result of Prediction Model

		Predicted by Model		
		Rise (1)	Drop (0)	
Actual	Rise (1)	TP = 40	FN = 19	59
	Drop (0)	FP = 6	TN = 84	90
		46	103	Total = 149

6.5 Data Analysis and Model Evaluation

The prediction model would be evaluated according to different criteria that are induced from the confusion matrix given by the SVM classifier. In the following section, we first evaluate the model based on the most common evaluation criterion, namely the accuracy, precision and recall and compare it with news random labeling. Then precision-recall F-measure, precision recall curve, and ROC curve would be discussed afterward.

Accuracy, Precision and Recall Evaluation

According to the confusion matrix, among the 149 pieces of news 59 of them are actually labeled as rise and 90 of them are actually labeled drop. From the 59 rise labeled news, the model predicts 40 of them correctly as rise and the remaining 19 are incorrectly labeled as drop. On the other hand, from the 90 drop labeled news, 84 of them are correctly labeled as drop and 6 of them are incorrectly labeled as rise. From these we calculate the model total accuracy, true positive rate (recall for rise category), true negative rate (recall for drop category), precision for rise, and precision for drop category.

1. Accuracy = (40+84)/149 = **83%**
2. True Positive Rate (Recall – Rise) = 40/59 = **67%**
3. True Negative Rate (Recall – Drop) = 84/90 = **93%**
4. Precision (Rise) = 40/46 = **87%**
5. Precision (Drop) = 84/ 103 = **81%**

The total accuracy of prediction model is equal to 0.83, which means 83% of time the model predicts correctly both the rise and the drop trends. But if you notice the true positive (67%) and true negative (93%) rates, you will realize that the model predicts the drop trend 1.38 times (93/67) better than the rise trend and the recall of drop category outperforms the recall of rise category. Inversely, precision of rise category do better than the precision of drop category. It means that among the total number of rise labeled by the model, more of them are actually labeled as rise (40 out of 46) but among the total number of drop labeled by the model, fewer of them (in compare with rise category) are actually labeled as drop.

For evaluating our prediction model, we compare it with news random labeling (labeling without the application of prediction model). In order to do so, we labeled the 149 news randomly as either 1(rise) or 0 (drop) by generating Bernoulli trial. The random labeling is then compared with the actual labels and the confusion matrix for news random labeling generated. The result of news random labeling is provided in Table 6.4.

Table 6.4: Confusion Matrix for News Random Labeling

		News Random Labeling		
		Rise (1)	Drop (0)	
Actual	Rise (1)	TP = 23	FN = 36	59
	Drop (0)	FP = 36	TN = 54	90
				Total = 149

Like the previous case, among the 149 news selected for prediction, 59 of them are actually labeled as rise and 90 of them are actually labeled as drop. Among the 59 actually rise labeled news, 23 of them are correctly predicted as rise, and 36 of them are incorrectly predicted as drop which indicates that most of the actually rise labeled news are predicted incorrectly as drop resulting in the true positive rate of less than 50% (TPR = 38%) which is no good at all. Among the 90 actually drop labeled news, 54 of them are correctly predicted as drop, and 36 of them are incorrectly predicted as rise.

1. Accuracy = (23+54)/149 = **51%**
2. True Positive Rate (Recall – Rise) = 23/59 = **38%**
3. True Negative Rate (Recall – Drop) = 54/90 = **60%**
4. Precision (Rise) = 23/59 = **38%**
5. Precision (Drop) = 54/90 = **60%**

Comparing the results obtained from labeling with classification model and labeling randomly, one can realize that using the classification model improves the prediction to a great extent. The total accuracy for random news labeling is equal to 51% and compared to accuracy of prediction model, we can conclude that prediction model predicts 1.62 times (83/51) better than the random prediction and the model improves the prediction 30% form 51% to 83%. The other measure (recall and precision) of random prediction is also much lower than the prediction model.

Precision-Recall F-measure

The performance measures (precision, recall) may be misleading when examined alone; we use another evaluation criterion that combines recall and precision called F-measure which is the weighted harmonic mean of precision and recall.

F1 (Rise Category-Prediction Model) = (2*precision*recall) / (precision + recall) = **0.75**
F1 (Drop Category-Prediction Model) = (2*precision*recall) / (precision + recall) = **0.86**

As was mentioned above, the recall of drop category was higher than the recall of rise category and inversely the precision of rise category was higher than the drop category and the decision on which category is performing better was a bit difficult, hence we combined the precision and recall measures using precision-recall F-measure to come up with a single number. Now we can surely say that drop category is predicted better than rise category. F-measure is also calculated for random labeling for both rise and drop categories to show that it is much less than the F-measure of prediction model.

F1 (Rise Category-Random) = (2*precision*recall) / (precision + recall) = **0.46**
F1 (Rise Category-Random) = (2*precision*recall) / (precision + recall) = **0.46**

Precision-Recall Curve

Another evaluation criterion that shows the relationship between precision and recall is characterized by precision-recall curve having recall on x-axis and precision on y-axis. We have illustrated the 4 different precision-recall curves in Figure 6.6. The pink and green curves are related to the prediction model, while green indicates the rise and pink indicates the drop category. The purple and blue curves are related to the random labeling, while purple relates to rise and blue relates to drop category. In general, as precision-recall curve moves toward the upper-left corner, the better performance is expected as both precision and recall of the prediction system is increasing. In the following section, we describe and analyze the concept of precision-recall curve and compare precision-recall curves of our prediction model with the random ones. We also compare the prediction model precision-recall curves of different categories.

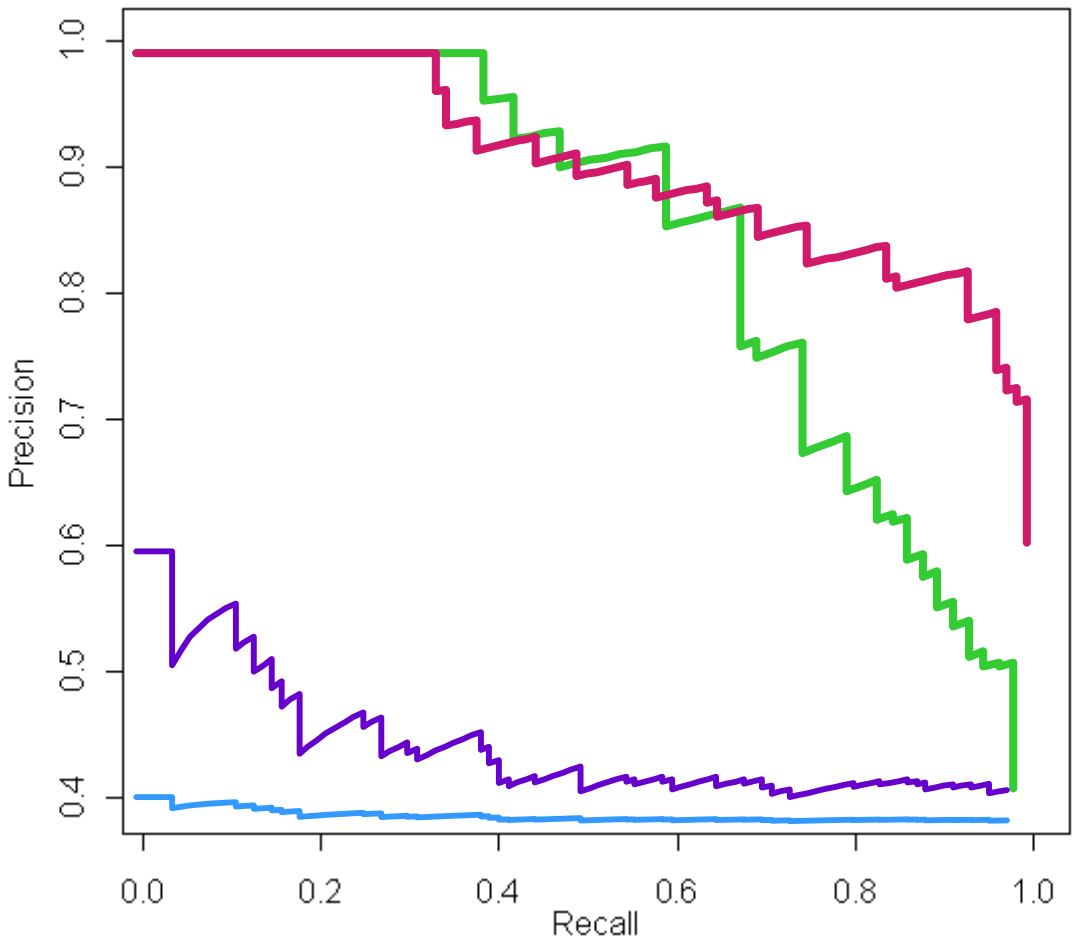


Figure 6.6: Precision-Recall Curve of Prediction Model vs. Random Precision-Recall

As was stated earlier, precision and recall may be misleading when examined alone. Precision-recall curve illustrate different value of precision for different recall rates. As it is shown in Figure 6.6, one can realize that the prediction model (pink and green curve) works much better than the random labeling (blue and purple curves) as it moves toward the upper-right corner. In prediction model we cannot say that the drop category (pink curve) outperform the rise category (green curve) at all times as there are some parts in the curve that the rise category has higher precision values relative to drop category for the same recall rates (0.3-0.6). The prediction model also outperforms the random labeling drawn in purple (drop category) and in blue (rise category). The maximum precision for random labeling in rise category is equal to starts from 0.39. In ransom labeling all the news are predicted as one and the precision would be the actual rise labeled divided by the total predicted as rise (59/149 = 0.39).

In order to depict the precision-recall curve, each of 149 news are assigned values which indicate the probabilities of being labeled as 1 or zero. The model sorts these probabilities and checks the actual and predicted labeling for each of them. As long as predicted labels are the same as the actual ones, the precision value is equal to one. For the recall rate between 0 and 0.35, the model predicts all the news correctly, afterwards there is a drop in precision value which indicates that the prediction model has incorrectly labeled the news as rise where in reality it is labeled drop. For all the points that there is a decrease in precision value, one drop labeled news is predicted as rise by the model.

Another way for visualizing the result of prediction model and compare it with random labeling is ROC curve having recall (TPR) on y-axis and false positive rate on x-axis. Figure 6.7 illustrates the ROC curve for prediction model and random labeling.

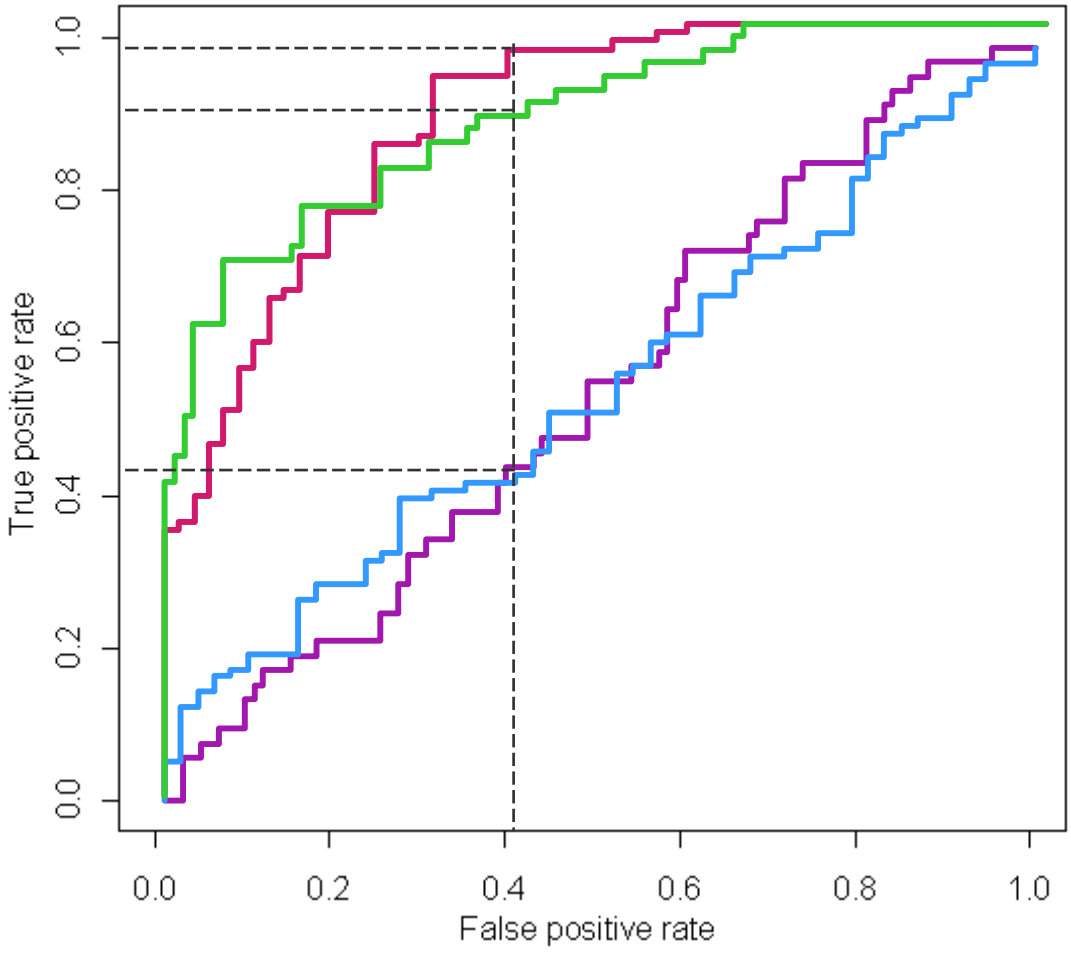


Figure 6.7: ROC Curve for Prediction Model vs. Random ROC Curve

ROC curve demonstrates the different proportions of the news correctly labeled to the news incorrectly labeled or the relationship between different true positive rates and true negative rates. As shown in Figure 6.7, ROC curve for the prediction model (green for rise and pink for drop) and the random labeling (blue for rise and purple for drop) have been depicted. The ideal case is the combination of higher true positive rate (TPR) and lower false positive rate (FPR) which can be obtained when ROC curve moves toward the upper-left corner. Comparing the prediction model ROC with random ROC, we can realize that the prediction model performs better than the random case. Consider the case when the FPR is equal to 40% means 40 percent of the news are labeled incorrectly. The associated TPR differs for prediction model and random labeling and even between the rise and drop category in prediction model. The associated TPR for 40% FPR is equal to almost 40 percent in random labeling, which means as much as the news are labeled correctly, with the same amount they are labeled incorrectly. But the associated TPR for prediction model is 90% for rise category and almost 100% for drop category which indicates that if 40% of the news are labeled incorrectly by the model, 90 percent of rise labeled news, and almost all the drop labeled news are predicted correctly.

From different evaluation criteria used in this study, we can conclude that our prediction model outperforms the random labeling and the model improves the accuracy of prediction from 51% in random labeling to 83%. In another expression, out of 10 news that are given to the prediction model to be labeled, 8 of them would be predicted correctly while in random labeling only 5 of them would be labeled correctly. We can claim that encouraging result is obtained for this experiment.

Chapter 7

Conclusion and Future Directions

7. Conclusion and Future Directions

In this chapter, the entire study would be reviewed briefly and the main results and concluding remarks are provided accordingly. The limitations and problems associated to the implementation of the research process would also be discussed. The managerial implications and the recommendations for future study would also be stated.

7.1 An Overview of Study

Stock markets have been studied over and over again to extract useful patterns and predict their movements. Mining textual documents and time series concurrently, such as predicting the movements of stock prices based on the contents of the news stories, is an emerging topic in data mining and text mining community. Stock price trend forecasting based solely on the technical and fundamental data analysis enjoys great popularity. But numeric time series data only contain the event and not the cause why it happened. Textual data such as news articles have richer information, hence exploiting textual information especially in addition to numeric time series data increases the quality of the input and improved predictions are expected from this kind of input rather than only numerical data. Information about company's report or breaking news stories can dramatically affect the share price of a security.

Financial analysts who invest in stock markets usually are not aware of the stock market behavior. They are facing the problem of stock trading as they do not know which stocks to buy and which to sell in order to gain more profits. All these users know that the progress of the stock market depends a lot on relevant news, but they do not know how to analyze all the news that appears on newspapers, magazines and other textual resources as the analysis of such amount of financial news and articles in order to extract useful knowledge exceeds their capabilities.

The main objective of this research is to predict the stock trend movement based on the contents of relevant news articles which can be accomplished by building a prediction model which is able to classify the news as either rise or drop. Making this prediction model is a binary classification problem which uses two types of data: past intraday price and past news articles. In order to make the model different data and text mining techniques should be applied to find the correlation between certain features found in these articles and changes in stock prices and the predictive model is learned through an appropriate classifier.

In order to make the prediction model, the research process should be implemented consists of different steps including data collection, data preprocessing, alignment, feature and document selection, document representation, classification and model evaluation. Each of these steps are coded either in R or Python programming languages and then are combined together to make the whole perdition package. As input we use real-time news articles and intra-day stock prices of Iran-khodro Company. A new statistical based piecewise segmentation algorithm is used to identify trends on the time series and news articles are preprocessed based on the Persian language. In order to label the news articles, they are aligned back to identified trends based on the Efficient Market Hypothesis. In order to filter news articles, the chi-square statistics is applied and selected documents are represented using vector space modeling and *tfidf* weighting scheme. Tfidf representation is given to SVM classifier and the prediction model is built accordingly. The model is then evaluated based on some evaluation criterion including accuracy, precision, recall, precision-recall F-measure, precision-recall curve, and ROC curve. The evaluation results are then compared with the news random labeling.

7.2 The Concluding Remark

Comparing the prediction model (machine learning model) accuracy (83%) with news random labeling accuracy (51%), we can conclude that our prediction model outperforms the random labeling. The prediction model will notify the up or down of the stock price movement when an upcoming pieces of news is released, and 83 percent of time can predict correctly. This can be very beneficial for individual and corporate investors, financial analysts, and users of financial news. With such a model they can foresee the future behavior and movement of stock prices; take correct actions immediately and act properly in their trading to gain more profit and prevent loss.

7.3 Limitations and Problems

Lack of Appropriate Data Mining Software

The most important limitation concerning the implementation of the research process is the lack of appropriate data mining software. As the research study is completely based on the application of data and text mining techniques, we needed some powerful tools and software to implement different steps in the research process. As there was no such data mining software at hand, we had to use R open source programming language and Python programming language to write the codes for the algorithms related to different steps in the research process and implement those algorithms in R and Python environment. Text preprocessing and classifier learning are two most important steps of the research process which require powerful tools to be implemented. Not having an automatic Persian text preprocessor, we have to code this process in Python environment Text preprocessing can not all be coded by a programming language as when dealing with text we are facing with thousands of words but having automatic software we can be sure that all these processes are implemented almost correctly and the rate of error would be much less. On the other hand, the classification process, which affects our prediction accuracy, is implemented in R which not a powerful tool is for Support Vector Machine Classifier and requires many manual parameter settings. Beside these, learning to write the coding and program in R and Python is quite a difficult and time-consuming task.

Inappropriate Database Management

One of the other problems regarding this research process is the lack of appropriate and adequate databases and data warehouses. As was mentioned earlier, data mining focuses on the computerized exploration of large amounts of data stored in databases and on the discovery of interesting patterns within them. Unfortunately most of the Iranian organizations are not aware of the importance of having adequate databases from which useful knowledge can be extracted. There was the difficulty of gathering past intraday stock prices of companies trading stocks in TSE. It was a time-consuming task as it took for months to find out where these data are actually stored.

Shortage of Online Time-Stamped News

Beside intraday stock prices of companies, we needed to gather enough number of online news from reliable online news providers and news websites so that data mining techniques and the research process can be applicable. On the 20 companies whose intraday stock prices are gathered, Iran-Khodro was the only company one whose number of news seemed almost enough for the purpose of this study and 1523 pieces of news gathered for the 2 consecutive years (1383 and 1384). But for the rest of companies hardly we could find two hundred to three hundred pieces of time-stamped news during these two years. In general there is the problem of having enough number of time-stamped news articles on the internet and if there are enough they are not time-stamped and the time of news release is not identified and hence not applicable for the purpose for this study. The other problem concerning news gathering about Iran-Khodro Company was the malfunctioning of the search engines provided by the online news providers. Some of them did not work at all, hence finding and searching manually the whole archive was not possible and those useful news could not be gathered. Some of them did not search two-parted words such as Iran-Khodro, and Khodro should be search alone and among the huge number of news on Khodro we had to save manually those related to Iran-Khodro Company. The prediction model could have performed better if more news could be gathered. The inadequate number of news and the malfunctioning of the search engines made this process a very difficult and time-consuming task.

7.4 Implications for Financial Investors

We suggest our prediction model to both individual and corporate investors and also to stock broker companies. The clear benefit of using such model is its profitability. If the investors take immediate actions on the stocks that they have at hand, they can beat the market, gain more profit or prevent loss. We have witnessed that many individual or corporate investors have been bankrupted as they have acted wrongly in their trading. The prediction model helps them to foresee the future behavior of stocks and take immediate actions upon them. The prediction model reduces the risk of loss up to 20 percent as 83 percent of time it predicts correctly. We recommend the prediction model to stock broker companies which will have several benefits for them. Tehran stock brokers do the trading for the individual investors and provide them with some recommendations of the status of different stocks. They suggest the investors which stocks to buy or which to sell just according to the status of market and there is no guarantee behind it. The accuracy of their predictions is the same as random labeling or a bit more because of their familiarity on financial markets status. But if they use the prediction model, what they suggest to their customers (individual investors) 80 percent of time is true and the customers would be satisfied as the stock broker has made them gain more profit. Keeping the customers satisfied has some benefits for the stock brokers. They can retain and increase their customers, improve the customer relationship management, and gain more profit.

7.5 Recommendation for Future Directions

Market Simulation

One of the best ways to evaluate the reliability of a prediction system is to conduct a market simulation which mimics the behaviors of investors in real-life data. One of the areas of further research is to conduct the market simulation on the proposed prediction model (shares are bought and sold based solely on the content of the news articles for a specified evaluation period) and on the Buy-and-Hold Test (stock are

bought at the beginning and sold at the end of the evaluation period. The rate of return for each simulation would be calculated and compared with each other.

Conducting Different Comparative Studies

In this research, we have made a prediction model by implementing and accomplishing a research process with identified methods and techniques. We would like to extend this research by applying other machine learning techniques and compare it with the machine learning techniques used in this study, namely the support vector machine. The comparative study not only lies in using different classification algorithm, but also in application of different techniques and approaches in the whole research process including the application of different feature selection criteria, other approaches in alignment process, and different methods of document representation. Application of different techniques in each of the research process will result in a new prediction model which would be the basis of a comparative study.

Evolving Trading Strategies

One of the other areas of further research is to evolve simple trading strategies after the model predicts the stock trend. It states that if the model predicts as up or down, which actions to take and how much stock to buy or sell or when to buy or sell. This is actually a complementary for the prediction model as after predictions it gives the afterward instructions to keep the investors in the best financial position.

Application of News Related to Automobile Industry

The types of news we have used for making the prediction model are those exactly related to the political, financial, production and other activities and policies of Iran-Khodro Company. We consider that other types of news related to the automobile industry as a whole and the news related to the other automobile competitors might have an effect on Iran-Khodro stock price movement. Hence we recommend making the prediction model applying all the types of news related to auto industry in general and the ones related to competitors and compare the results with the current prediction model.

Reference:

Aas, K., and Eikvil, L., 1999. *Text Categorization: A Survey*. Technical Report NR 941. Oslo, Norway: Norwegian Computing Center (NR).

Achelis, S.B., 1995. *Analysis from A to Z.* 2nd ed. Chicago: Irwin Professional Publishing.

Achlioptas, D., 2001. Database Friendly Random Projections. In *Proceedings of the 20th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS), (Santa Barbara, CA, May 21-23, 2001)*. New York: ACM Press, 2001, pp.274-281.

Agrawal, C.C., and Yu, P.S., 2000. Finding Generalized Projected Clusters in High Dimensional Spaces. In W. Chen, J.F. Naughton, and P.A. Bernstein eds. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data (Dallas, Texas, May 15-18, 2000)*. New York (NY): ACM Press, 2000, pp.70-81.

Agrawal, R., Lin, K., Sawhney, H.S., and Shim, K., 1995. Fast Similarity Search in the Presence of Noise, Scaling, and Translation in Time Series Databases. In U. Dayal, P.M. Gray, and S. Nishio, eds. *Proceedings of the 21st International Conference on Very Large Data Bases (VLDB), (Zurich, Switzerland, September 11-15, 1995)*. San Francisco, California: Morgan Kaufmann Publishers Inc., 1995, pp. 490-501.

Albrecht, R., and Merkl, D., 1998. Knowledge Discovery in Literature Data Bases. In U. Grothkopf, H. Andernach, S. Stevens-Rayburn, and M. Gomez eds. *The 3rd Conference on Library and Information Services in Astronomy (Tenerife, Spain, April 21-24, 1998), ASP (Astronomical Society of the Pacific) Conference Series,* vol.153. pp.93-101.

Allen, F., and Karjalainen, R., 1995. Using Genetic Algorithms to Find Technical Trading Rules. *Journal of Financial Economics*, 51(2), pp. 245-271.

Anghelescu, A.V., and Muchnik, I.B., 2003. Combinatorial PCA and SVM Methods for Feature Selection in Learning Classifications: Applications to Text Categorization. In *IEEE International Conference on Integration of Knowledge Intensive Multi-Agent Systems (KIMAS), (Boston, Miami, October 01-03, 2003)*. IEEE Press, 2003, pp.491-496.

Apte, C., Damerau, R., and Weiss, S.M., 1994. Automated Learning of Decision Rules for Text Categorization. *ACM Transactions on Information Systems*, 12(3), pp.233-251.

Baker, L., and McCallum, A., 1998. Distributional Clustering of Words for Text Classification. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Melbourne, Australia, August 24-28, 1998)*. New York (NY): ACM Press, 1998, pp.96-103.

Basili, R., Moschitti, A., and Pazienza, M.T., 2001. A Hybrid Approach to Optimize Feature Selection Process in Text Classification. In F. Esposito ed. *Advances in Artificial*

Intelligence, Proceedings of the 7th Congress of the Italian Association for Artificial Intelligence (Bari, Italy, September 25-28, 2001), Lecture Notes In Computer Science. Heidelberg, Berlin: Springer-Verlag, 2001, vol.2175, pp.320-326.

Basu, A., Watters, C., and Shepherd, M., 2003. Support Vector Machines for Text Categorization. In *Proceedings of the 36th Annual Hawaii International Conference on System Sciences (HICSS), (Big Island, Hawaii, January 06-09, 2003)*. Washington, DC: IEEE Computer Society, 2003, 4(4), pp.103-109.

Battiti, R., 1994. Using Mutual Information for Selecting Features in Supervised Neural Net Learning. *IEEE Transactions on Neural Networks*, 5(4), pp.537-550.

Berry, T.D., Howe, K.M., 1994. Public Information Arrival. *Journal of Finance,* 49(4), pp.1331–1346.

Biggs, M., 2000. Enterprise Toolbox: Resurgent Text-mining Technology Can Greatly Increase Your Firm's 'Intelligence' Factor. *InfoWorld*, 11(2). [Online]
Available from: http://www.infoworld.com/articles/op/xml/00/01/10/000110opbiggs.html

Bingham, E., and Mannila, H., 2001. Random Projection in Dimensionality Reduction: Applications to Image and Text Data. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), (San Francisco, California, August 26-29, 2001)*. New York: ACM Press, 2001, pp. 245-250.

Blum, A.L., and Langley, P., 1997. Selection of Relevant Features and Examples in Machine Learning. *Artificial Intelligence*, *1(2)*, 245-271

Bong, C.H., Narayanan, K. and Wong, T.K., 2005. An Examination of Feature Selection Frameworks in Text Categorization. In G.G. Lee, A. Yamada, H. Meng, and S.H. Myaeng eds. *Information Retrieval Technology, Proceedings of the 2nd Asia Information Retrieval Symposium (AIRS), (Jeju Island, Korea, October 13-15, 2005), Lecture Notes in Computer Science*. Heidelberg, Berlin: Springer-Verlag, 2005, vol.3689, pp.558-564.

Borges, G.A., and Aldon, M.J., 2000. A Split-and-Merge Segmentation Algorithm for Line Extraction in 2-D Range Images. In *Proceedings of the 15th International Conference on Pattern Recognition (ICPR), (Barcelona, Spain, September 03-08, 2000)*. Washington, DC: IEEE Computer Society, vol.1, pp.1441-1444.

Bouchard, D. (n.d.) Automated Time Series Segmentation for Human Motion Analysis. Philadelphia: Center for Human Modeling and Simulation, University of Pennsylvania. [Online]. Available from:
hms.upenn.edu/RIVET/SupportingDocs/AutomatedTimeSeriesSegmentation.pdf

Boulis, C., and Ostendorf, M., 2005. Text Classification by Augmenting the Bag-of-Words Representation with Redundancy Compensated Bi-grams. In *Workshop on Feature Selection in Data Mining (FSDM) at the SIAM International Conference on Data*

Mining (Newport Beach, California, April 21-23, 2005). Workshop Proceedings [Online]. Available from:http://enpub.eas.asu.edu/workshop/FSDM05-Proceedings.pdf

Brucher, H., Knolmayer, G., and Mittermayer, M.A., 2002. Document Classification Methods for Organizing Explicit Knowledge. In *Proceedings of the 3rd European Conference on Organizational Knowledge, Learning and Capabilities (OKLC), (Athens, Greece, April 05-06, 2002)*. Proceedings Available at ALBA. [Online].
Available from: http://www.alba.edu.gr/OKLC2002/Proceedings/track7.html

Buckley, C., Salton, G., and Allan, J., 1994. The Effect of Adding Relevance Information in a Relevance Feedback Environment. In W.B. Croft and C.J. van Rijsbergen eds. *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Dublin, Ireland, July 03-06, 1994)*. New York (NY): Springer-Verlag, 1994, pp.292-300.

Chakrabarti, S., 2003. Mining the Web. New York: Morgan Kaufmann Publishers.

Chan Y., Chui, A.C.W., and Kwok, C.C.Y., 2001. The Impact of Salient Political and Economic News on the Trading Activity. *Pacific-Basin Finance Journal*,9(3),pp.195-217.

Chan, Y., and John-Wei, K.C., 1996. Political Risk and Stock Price Volatility: The Case of Hong Kong. *Pacific-Basin Finance Journal*, 4(2-3), pp.259-275.

Chen, H., Hsu, P., Orwig, R., Hoopes, L., and Nunamaker, J.F., 1994. Automatic Concept Classification of Text from Electronic Meetings. *Communications of ACM*, 37(10), pp.56-73.

Chung, F., Fu, T., Luk, R., and Ng, V., 2002. Evolutionary Time Series Segmentation for Stock Data Mining. In *Proceedings of IEEE International Conference on Data Mining (Maebashi, Japan, Dec. 09-12, 2002*. Washington: IEEE Computer Society, , pp.83-91.

Cohen, W.W., and Singer, Y., 1996. Context-Sensitive Learning Methods for Text Categorization. *ACM Transactions on Information Systems (TOIS)*, 17(2), pp. 141-173.

Cooper, D.R., and Schindler, P.S., 2003. *Business Research Methods*. 8th ed. New York: McGraw-Hill.

Cortes, C., and Vapnik, V., 1995. Support Vector Networks. *Machine Learning*, 20(3), pp. 273-297

Creecy, R.H., Masand, B.M., Smith, S.J., and Waltz, D.L., 1992. Trading MIPS and Memeory for Knowledge Engineering: Classifying Census Returns on the Connection Machine. *Communication of the ACM,* 35(8), pp.48-64.

Cristianini, N., and Shawe-Taylor, J., 2002. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge: Cambridge University Press.

Das, G., Lin, K.I., and Mannila, H., Renganathan, G., Smyth, P., 1998. Rule Discovery from Time Series. In R. Agrawal, P.E. Stolorz, G. Piatetsky-Shapiro eds. *Proceedings of the Fourth ACM International Conference on Knowledge Discovery and Data Mining(KDD), (New York, August 27-31, 1998)*. New York: AAAI Press, 1998, pp.16-22.

Dash, M., and Liu, H., 1997. Feature Selection for Classification. *International Journal of Intelligent Data analysis*, Elsevier, 1(3), pp.131-156.

Dash, M., & Liu, H. 2000. Feature Selection for Clustering. In T. Terano, H. Liu, and A.L.P. Chen eds. *Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Current Issues and New Application (Kyoto, Japan, April 18-20, 2000), Lecture Notes in Computer Science*. London: Springer-Verlag, 2000, vol.1805, pp.110-121.

Davis, J., and Goadrich, M., 2006. The Relationship between Precision-Recall and ROC Curves. In W.W. Cohen and A. Moore eds. *Proceedings of the 23rd International Conference on Machine Learning (ICML), (Pittsburgh, Pennsylvania, June 25-29, 2006)*. New York (NY): ACM Press, 2006, vol. 148, pp.233-240.

Debole, F., and Sebastiani, F., 2002. *Supervised Term Weighting for Automated Text Categorization*. Technical Report 2002-TR-08. Pisa, Italy: Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche.

Debole, F., and Sebastiani F., 2003. Supervised Term Weighting for Automated Text Categorization. In *Proceedings of the 18th ACM Symposium on Applied Computing (SAC), (Melbourne, Florida, March 09-12, 2003)*. New York (NY): ACM Press, 2003, pp.784-788.

Deegalla, S., and Bostrom, H., 2006. Reducing High-Dimensional Data by Principal Component Analysis vs. Random Projection for Nearest Neighbor Classification. In *Proceedings of the 5th International Conference on Machine Learning and Applications (ICMLA), (Orlando, Florida, December 14-16, 2006)*. IEEE, 2006, pp.245-250.

Deerwester, S., Dumais, S.T., Landauer T.K., Furnas, G.W., and Harshman, R.A., 1990. Indexing by Latent Semantic Analysis. *Journal of the Society for Information Science*, 41(6), pp.391-407.

Doan, S., and Horiguchi, S., 2004a. An Agent-based Approach to Feature Selection in Text Categorization. In S.C. Mukhopadhyay and G. Sen Gupta eds. *Proceedings of the 2nd International Conference on Autonomous Robot and Agent (ICARA), (Palmerston North, New Zealand, Dec. 13-15, 2004)*. New Zealand: Massey University, pp.262-366.

Doan, S., and Horiguchi, S., 2004b. An Efficient Feature Selection Using Multi-Criteria in Text Categorization. In *Proceedings of the 4th International Conference on Hybrid Intelligent Systems (HIS), (Kitakyushu, Japan, December 05-08, 2004)*. Washington, DC:

IEEE Computer Society, 2004, pp.86-91.
Dorre, J., Gerstl, P., and Seiffert, R., 1999. Text Mining: Finding Nuggets in Mountains of Textual Data. In U. Fayyad, S. Chaudhuri, and D. Madigan ed. *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (San Diego, CA, Aug. 15 - 18, 1999).* New York: ACM Press, 1999, pp.398-401.

Dumais, S., and Chen, H., 2000. Hierarchical Classification of Web Content. In E. Yannakoudakis, N.J. Belkin, M.K. Leong, and P. Ingwersen eds. *Proceedings of the 23rd Annual International SIGIR Conference on Research and Development in Information Retrieval (Athens, Greece, July 24-28, 2000).* New York: ACM Press, pp.256-263.

Dumais, S., Platt, J., Heckerman, D., and Sahami, M., 1998. Inductive Learning Algorithms and Representations for Text Categorization. In G. Gardarin, J.C. French, N. Pissinou, K. Makki, and L. Bouganim eds. *Proceedings of the 7th ACM International Conference on Information and Knowledge Management (CIKM) (Bethesda, Maryland, November 02-07, 1998).* New York (NY): ACM Press, 1998, pp.148-155.

Dunning, T., 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics,* 19(1), pp.61-74.

Even-Zohar, Y., 2002. Introduction to Text Mining, Part II. *Presentation 2 of a 3-part Series Given at SC 2002 by the Automated Learning Group (ALG).* National Center for Supercomputing Applications, University of Illinois. [Online]
Available from: algdocs.ncsa.uiuc.edu/PR-20021116-2.ppt

Everitt, B.S., 1993. *Cluster Analysis*. 3rd ed. New York (NY): John Wiley and Sons, Inc., London: Edvard Arnold, New York: Halsted Press.

Eyheramendy, S., and Madigan, D., 2005. A Novel Feature Selection for Text Categorization. In *Workshop on Feature Selection in Data Mining (FSDM) at the SIAM International Conference on Data Mining (Newport Beach, California, April 21-23, 2005).* Workshop proceedings are available [Online].
Available from: http://enpub.eas.asu.edu/workshop/FSDM05-Proceedings.pdf

Faloutsos, C., Ranganathan, M., and Manolopoulos, Y., 1994. Fast Subsequence Matching in Time Series Databases. In R.T. Snodgrass and M. Winslett eds. *Proceedings of the 1994 ACM SIGMOD International Conference on Management of Data (SIGMOD) (Minneapolis, Minnesota, May 24-27, 1994*). New York: ACM Press, 2001, pp.419-429.

Fama, E.F., 1964. *The Distribution of the Daily Differences of the Logarithms of Stock Prices*. Unpublished Ph.D Dissertation. Chicago: University of Chicago.

Fama, E.F., 1970. Efficient Capital Markets: A Review of Theory and Empirical Work. *Papers and Proceedings of the Twenty-Eighth Annual Meeting (New York, December 28-30, 1969) of American Finance Association, Journal of Finance,* 25(2), pp.383-417.

Fama, E.F., 1991. Efficient Capital Markets: II. *Journal of Finance*, 46(5), pp.1575-1617.
Fawcett, T., and Provost, F., 1999. Activity Monitoring: Noticing Interesting Changes in Behavior. In S. Chaudhuri, D. Madigan, and U. Fayyad eds. *Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD, San Diego, California, August 15-18, 1999)*. New York: ACM Press, 1999, pp.53-62.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P., 1996a. The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Communication of ACM*. New York: ACM Press, 39(11), pp.27-34.

Fayyad U.M., Piatetsky-Shapiro, G., and Smyth, P., 1996b. From Data Mining to Knowledge Discovery: An Overview. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy eds. *Advances in Knowledge Discovery and Data Mining*. Cambridge, Massachusetts: AAAI / MIT Press, 1996, pp.1-34. Also in *AI Magazine*, 17(3), pp.37-54.

Fern, X.Z., and Brodley, C.E., 2003. Random Projection for High Dimensional Data Clustering: A Cluster Ensemble Approach. In T. Fawcett and N. Mishra eds. *Proceedings of the 20th International Conference on Machine Learning (ICML), (Washington, DC, August 21-24, 2003)*. Menlo Park, California: AAAI Press, 2003, pp.186-193.

Fodor, I.K., 2002. A Survey of Dimension Reduction Techniques. [Online]. Livermore, California: Center for Applied Scientific, Lawrence Livermore National Laboratory. Available from: www.llnl.gov/CASC/sapphire/pubs/148494.pdf

Forman, G., 2002. Choose Your Words Carefully: An Empirical Study of Feature Selection Metrics for Text Classification. In T. Elomaa, H. Mannila, and H. Toivonen eds. *Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD), (Helsinki, Finland, August 19-23, 2002), Lecture Notes in Artificial Intelligence*. Heidelberg, Berlin: Springer-Verlag, 2002, vol.2431, pp150-162.

Forman, G., 2003. An Extensive Empirical Study of Feature Selection Metrics for Text Classification. *Journal of Machine Learning Research,* vol. 3, pp.1289-1305.

Fradkin, D., and Madigan, D., 2003. Experiments with Random Projections for Machine Learning. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), (Washington, DC, August 24-27, 2003)*. New York (NY): ACM Press, 2003, pp. 517-522.

Fragos, K., Maistros, Y., and Skourlas, C., 2005. A Weighted Maximum Entropy Language Model for Text Classification. In B. Sharp ed. *Proceedings of the 2nd International Workshop on Natural Language Understanding and Cognitive Science (NLUCS) (Miami, FL, May 24, 2005)*. Miami, Florida: INSTICC Press, 2005, pp.55-67.

Frawley, W.J., Piatetsky-Shapiro, G., and Matheus, C.J., 1991. Knowledge Discovery in Databases: An Overview. In G. Piatetsky-Shapiro and W.J. Frawley eds. *Knowledge*

Discovery in Databases. Menlo Park, California (CA): AAAI/MIT Press, 1991, pp.1-30. Reprinted in Fall 1992 in *AI Magazine*, 13(3), pp.57-70.

Fuhr, N., Hartmanna, S., Knorz, G., Lusting, G., Schwantner, M., and Tzeras, K., 1991. Air/X – a Rule-Based Multistage Indexing Systems for Large Subject Fields. In A. Lichnerowicz ed. *Proceedings of the 3rd RIAO Conference (Barcelona, Spain, April 02-05, 1991)*. Amsterdam: Elsevier Science Publishers, 1991, pp.606-623.

Fukumoto, F., and Suzuki, Y., 2001. Learning Lexical Representation for Text Categorization. *Proceedings of the 2nd NAACL (North American Chapter of the Association for Computational Linguistics) Workshop on Wordnet and Other Lexical Resources: Applications, Extensions and Customizations (Pittsburgh, Pennsylvania, June 03-04, 2001)*.

Fung G.P.C., Yu, J.X., and Lam, W., 2002. News Sensitive Stock Trend Prediction. In M.S. Chen, P.S. Yu, and B. Liu, eds. *Proceedings of the 6th Pacific-Asia Conference (PAKDD) on Advances in Knowledge Discovery and Data Mining (Taipei, Taiwan, May 06-08, 2002), Lecture Notes in Computer Science*. Heidelberg, Berlin: Springer-Verlag, 2002, Vol.2336, pp.481-493.

Fung G.P.C., Yu, J.X., and Lam, W., 2003. Stock Prediction: Integrating Text Mining Approach Using Real-time News. In *Proceedings of the 7th IEEE International Conference on Computational Intelligence for Financial Engineering (CIFEr) (Hong Kong, China, March 20-23, 2003)*, IEEE Press, pp.395–402.

Fung G.P.C., Yu, J.X., and Lu, H., 2005. The Predicting Power of Textual Information on Financial Markets. *IEEE Intelligent Informatics Bulletin*, 5(1), pp.1-10.

Galavotti, L., Sebastiani, F., and Simi, M., 2000. Experiments on the Use of Feature Selection and Negative Evidence in Automated Text Categorization. In J.L. Borbinha and T. Baker eds. *Research and Advanced Technology, Proceedings of the 4th European Conference on Digital Libraries, (Lisbon, Portugal, September 18-20, 2000), Lecture Notes in Computer Science*. Heidelberg, Berlin: Springer-Verlag, vol.1923, pp.59-68.

Ge, X., 1998. Pattern Matching in Financial Time Series Data. In *Final Project Report for ICS 278*. Irvin: Department of Information and Computer Science, University of California. [Online]. Available from: http://citeseer.ist.psu.edu/334311.html

Gidofalvi, G., 2001. Using News Articles to Predict Stock Price Movements. San Diego: Department of Computer Science and Engineering, University of California. [Online]
Available from: http://www.cs.aau.dk/~gyg/docs/financial-prediction.pdf

Gidofalvi, G., and Elkan, C., 2003. *Using News Articles to Predict Stock Price Movements.* Technical Report. San Diego: Department of Computer Science and Engineering, University of California. [Online]
Available from: http://www.cs.aau.dk/~gyg/docs/financial-prediction-TR.pdf

Gilad-Bachrach, R., Navot, A., and Tishby, N., 2004. Margin Based Feature Selection - Theory and Algorithms. In C.E. Brodley ed. *Proceedings of the 21st International Conference on Machine Learning (ICML), (Banff, Alberta, Canada, July 04-08, 2004)*. New York (NY): ACM Press, 2004, vol.69, pp.43-50.

Goadrich, M., Oliphant, L., and Shavlik, J., 2006. Gleaner: Creating Ensembles of First-Order Clauses to Improve Recall-Precision Curves. *Journal of Machine Learning*, 64(1-3), pp.231-261.

Goutte, C., and Gaussier, E., 2005. A Probabilistic Interpretation of Precision, Recall, and *F*-Score, with Implication for Evaluation. In D.E. Losada and J.M. Fernández-Luna eds. *Advances in Information Retrieval, Proceedings of the 27th European Conference on Information Retrieval Research (ECIR) (Santiago de Compostela, Spain, March 21-23, 2005), Lecture Notes in Computer Science*. Heidelberg, Berlin: Springer-Verlag, 2005, vol. 3408, pp.345-359.

Grobelnik, M., Mladenic, D., and Milic-Frayling, N., 2000. Text Mining as Integration of Several Related Research Area: Reports on KDD 2000 Workshop on Text Mining. *ACM SIGKDD Explorations Newsletter*, 2 (2), pp.99-102.

Guo, G., Wang, H., and Bell, D.A., 2002. Data Reduction and Noise Filtering for Predicting Times Series. In X. Meng, J. Su, and Y. Wang eds. *Proceedings of the Third International Conference on Advances in Web-Age Information Management (Beijing, China, August 11-13*, 2002), *Lecture Notes In Computer Science*. London: Springer-Verlag, 2002, Vol.2419, pp.421-429.

Guyon, I., and Elisseeff, A., 2003. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, vol.3, pp.1157-1182.

Han, J., Dong, G., and Yin, Y., 1999. Efficient Mining of Partial Periodic Patterns in Time Series Database. In *Proceedings of the Fifteenth International Conference on Data Engineering (ICDE) (Sydney, Australia, March 23-26, 1999)*. Washington, DC: IEEE Computer Society, 1999, pp.106-116.

Hardin, D.P., Tsamardinos I., and Aliferis, C.F., 2004. A Theoretical Characterization of Linear SVM-based Feature Selection. In C.E. Brodley ed. *Proceedings of the 21st International Conference on Machine Learning (ICML), (Banff, Alberta, Canada, July 04-08, 2004)*. New York (NY): ACM Press, 2004, vol.69, pp.48-55.

Hariharan, G., 2004. *News Mining Agent for Automated Stock Trading*. Unpublished Master's Thesis. Austin: University of Texas.

Hearst, M.A., 1997. Text Data Mining: Issues, Techniques, and the Relationship to Information Access. *Presentation Notes for UW/MS Workshop on Data Mining*. [Online]. Available from: www.ischool.berkeley.edu/~hearst/talks/dm-talk

Hearst, M.A., 1999. Untangle Text Data Mining. *In Proceedings of the 37th conference on Association for Computational Linguistics on Computational Linguistics (Annual Meeting of ACL, College Park, Maryland, June 20-26, 1999)*. Morristown, New Jersey (NJ): Association for Computational Linguistics, 1999, pp.3-10.

Hearst, M.A., 2003. What is Text Mining? Berkley: The School of Information Management and Systems (SIMS), University of California. [Online]
Available from: www.sims.berkeley.edu/~hearst/text-mining.html [cited in April 2006]

Hearst, M.A., Schoelkopf, B., Dumais, S., Osuna, E., and Platt, J., 1998. Trends and Controversies - Support Vector Machines. *IEEE Intelligent Systems*, 13(4), pp.18-28.

Hellstrom, T., and Holmstrom, K., 1998. *Predicting the Stock Market*. Technical Report Series IMa-TOM-1997-07. Sweden: Center of Mathematical Modeling (CMM), Department of Mathematics and Physics, Malardalen University.

Hirshleifer, D., and Shumway T., 2003. Good Day Sunshine: Stock Returns and the Weather. *Journal of Finance*, 58(3), pp.1009-1032.

How, B.C., and Narayanan, K., 2004. An Empirical Study of Feature Selection for Text Categorization Based on Term Weightage. In *Proceedings of IEEE/WIC/ACM International Joint Conference on the Web Intelligence (WI), (Beijing, China, September 20-24, 2004)*. Washington, DC: IEEE Computer Society, 2004, pp.599-602.

Jain, A.K., Duin, P.W., and Jianchang, M., 2000. Statistical Pattern Recognition: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*,22(1), pp.4-37.

Jenkins, C., Jackson, M., Burden, P., and Wallis, J. 1999. Automatic RDF Metadata Generation for Resource Discovery. *Computer Networks: The International Journal of Computer and Telecommunications Networking*, 31(11-16), pp.1305-1320.

Joachims, T., 1997. A probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. In D.H. Fisher ed. *Proceedings of the 14th International Conference on Machine Learning (ICML), (Nashville, Tennessee, July 08-12, 1997)*. San Francisco, California: Morgan Kaufmann Publishers Inc., 1997, pp.143-151.

Joachims, T., 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In C. N\'edellec and C. Rouveirol eds. P*roceedings of the 10th European Conference on Machine Learning* (*ECML), Application of Machine Learning and Data mining in Finance (Chemnitz, Germany, April 21-24, 1998), Lecture Notes in Computer Science*. Heidelberg, Berlin: Springer Verlag, 1998, 1398(2), pp.137-142.

Joachims, T., 2002. *Learning to Classify Text Using Support Vector Machines: Methods, Theory, and Algorithms*. Norwell, Massachusetts: Kluwer Academic Publishers.

John, G.H., Kohavi, R., and Pfleger, K., 1994. Irrelevant Features and the Subset Selection Problem. In W.W. Cohen and H. Hirsh eds. *Proceedings of the 11th International Conference in Machine Learning (ICML), (New Brunswick, New Jersey, July 10-13, 1994)*. San Francisco, CA: Morgan Kaufmann Publishers, 1994, pp.121-129.

Johnson, W.B., and Lindenstrauss, J., 1984. Extensions of Lipshitz Mapping into Hilbert Space. In R. Beals ed. *Contemporary Mathematics, Proceedings of Conference in Modern Analysis and Probability*. Providence, Road Island: American Mathematical Society Publishers, 1984, vol.26, pp. 189-206.

Jolliffe, I.T., 1986. *Principal Component Analysis*. New York (NY): Springer-Verlag, Series in Statistics.

Kai, O.Y., Jia, W., Zhou, P., and Meng, X., 1999. A New Approach to Transforming Time Series into Symbolic Sequences. In *Proceedings of the First Joint BMES/EMBS Conference Serving Humanity Advancing Technology (Atlanta, Georgia, October 13-16, 1999)*. Piscataway, New Jersey (NJ): IEEE Computer Society Press, Vol.2, on Page 974.

Karanikas, H., and Theodoulidis, B., 2002. Knowledge Discovery in Text and Text Mining Software. Centre for Research in Information Management (CRIM), UMIST University, UK. [Online], Available from: www.crim.co.umist.ac.uk/parmenides/internal/docs/Karanikas_NLDB2002%20.pdf

Kaski, S., 1998. Dimensionality reduction by Random Mapping: Fast Similarity Computation for Clustering. In *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN): IEEE World Congress on Computational Intelligence, (Anchorage, Alaska, May 04-09, 1998)*. Piscataway, New Jersey (NJ): IEEE Computational Intelligence Society, 1998, vol.1, pp.413-418.

Kaufman, L., and Rousseeuw, P.J., 1990. *Finding Groups in Data – An Introduction to Cluster Analysis*. New York (NY): John Wiley and Sons Inc. (Series in Applied Probability and Statistics)

Keerthi, S.S., 2005. Generalized LARS as an Effective Feature Selection Tool for Text Classification with SVMs. In L. De Raedt and S. Wrobel eds. *Proceedings of the 22nd International Conference on Machine Learning (ICML), (Bonn, Germany, August 07-11, 2005)*. New York (NY): ACM Press, 2005, vol.119, pp.417-424.

Keogh, E.J., 1997. A Fast and Robust Method for Pattern Matching in Time Series Databases. In *Proceedings of the 9th International Conference on Tools with Artificial Intelligence (ICTAI), (Newport Beach, CA, November 03-08, 1997)*. Washington, DC: IEEE Computer Society, 1997, pp.578-584.

Keogh, E.J, Chakrabarti, K., Pazzani, M., and Mehrotra, S., 2001b. Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases. *Knowledge and Information Systems*. London: Springer-Verlag, 3(3), pp.263-286.

Keogh, E.J., Chu, S., Hart, D., and Pazzani, M.J., 2001a. An Online Algorithm for Segmenting Time Series. In N. Cercone, T.Y. Lin, and X. Wu, eds. *Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM), (San Jose, CA, November 29-December 02, 2001)*. Washington, DC: IEEE Computer Society, 2001, pp.289-296.

Keogh, E.J., and Kasetty, S., 2002. On the need for Time Series Data Mining Benchmarks: a Survey and Empirical Demonstration. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) (Alberta, Canada, July 23 - 26, 2002)*. New York (NY): ACM Press, 2002, pp.102-111.

Keogh, E.J., and Pazzani, M.J., 1998. An Enhanced Representation of Time Series which Allows Fast and Accurate Classification, Clustering and Relevance Feedback. In R. Agrawal, P.E. Stolorz, G. Piatetsky-Shapiro eds. *Proceedings of Fourth ACM International Conference on Knowledge Discovery and Data Mining (KDD), (New York, August 27-31, 1998)*. New York (NY): AAAI Press, 1998, pp.239-243.

Keogh, E.J., and Pazzani, M.J., 1999. Relevance Feedback Retrieval of Time Series Data. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Berkeley, California, August 15-19, 1999)*. New York (NY): ACM Press, 1999, pp.183-190.

Keogh, E.J., and Smyth, P., 1997. A Probabilistic Approach to Fast Pattern Matching in Time Series Databases. In D. Heckerman, H. Mannila, D. Pregibon eds. *Proceedings of 3rd International Conference on Knowledge Discovery and Data Mining* (*KDD), (Newport Beach, CA, August 14-17, 1997)*. Menlo Park, CA: AAAI Press,1997, pp.24-30.

Khare, R., Pathak, N., Gupta, S.K., and Sohi, S., 2004. Stock Broker P – Sentiment Extraction for the Stock Market. In A. Zanasi, N.F.F. Ebecken, and C.A. Brebbia eds. *Data Mining V, Proceedings of the Fifth International Conference on Data Mining, Text Mining and Their Business Applications (Malaga, Spain, Sep. 15-17, 2004)*. Southampton, Boston: WIT Press, 2004, Vol.33, pp.43-52.

Klein, F.C., and Prestbo, J.A., 1974. *News and the Market*. Chicago: Henry Regnery.

Klibanoff, P., Lamont, O., and Wizman, T.A., 1998. Investor Reaction to Salient News in Closed-end Country Funds. *Journal of Finance*, 53(2), pp.673-699.

Kohavi, R., and John, G.H., 1997. Wrappers for Feature Subset Selection. *Artificial Intelligence*, 97(1-2), pp.273-324.

Koller, D., and Sahami, M., 1996. Toward Optimal Feature Selection. In L. Saitta ed. *Proceedings of the 13th International Conference on Machine Learning (ICML), (Bari, Italy, Jul. 3-6, 1996)*. California: Morgan Kaufmann Publishers Inc., 1996, pp.284-292.

Kroeze, J.H., Matthee, M.C., and Bothma, T.J., 2003. Differentiating Data and Text Mining Terminology. In J. Eloff, A. Engelbrecht, P. Kotzé, and M. Eloff, eds. *Proceedings of the ACM 2003 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists (SAICSIT) on Enablement Through Technology (Johannesburg, Sep. 17-19, 2003*). South Africa: South African Institute for Computer Scientists and Information Technologists, 2003, vol. 47, pp.93-101.

Kroha, P., Baeza-Yates, R., 2004. *Classification of Stock Exchange News*. Technical Report. Department of Computer Science, Engineering School, Universidad de Chile.

Kwok, J.T., 1998. Automated Text Categorization Using Support Vector Machine. In S. Usui and T. Omori eds. *Proceedings of the 5th International Conference on Neural Information Processing (ICONIP), (Kitakyushu, Japan, October 21-23, 1998)*. San Francisco, California: IOA (Institute of Aging) Press, 1998, vol.1, pp.347–351.

Kwon, O.W., and Lee, J.H., 2000. Web Page Classification Based on K-Nearest Neighbor Approach. In K.F. Wong, D.L. Lee, and J.H. Lee eds. *Proceedings of the 5th International Workshop on Information Retrieval with Asian Languages (IRAL), (Hong Kong, China, September 30-October 1, 2000)*. New York: ACM Press, 2000, pp.9-15.

Lam, W., Low, K.F. and Ho, C.Y., 1997. Using a Bayesian Network Induction Approach for Text Categorization. In *Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI), (Nagoya, Japan, August 23-29, 1997)*. San Francisco, California: Morgan Kaufmann Publishers Inc., 997, vol.2, pp.745-750.

Landgrebe, T.C., Paclik, P., and Duin, R.P., 2006. Precision-Recall Operating Characteristic (P-ROC) Curves in Imprecise Environments. In B. Werner ed. *Proceedings of the 18th International Conference on Pattern Recognition (ICPR), (Hong Kong, China, August 20-24, 2006)*. Washington, DC: IEEE Computer Society, 2006, vol.4, Track 2, pp.123-127.

Langley, P., 1994. Selection of Relevant Features in Machine Learning. In: *Proceedings of the AAAI Fall Symposium on Relevance (New Orleans, Louisiana, November 04-06, 1994*). Menlo Park, California: AAAI Press, 1994, pp.1-5.

Larkey, L.S., 1998. Some Issues in the Automatic Classification of U.S. Patents. In *Workshop on Learning for Text Categorization, at the 15th National Conference on Artificial Intelligence (Madison, WI, July 26-30, 1998)*. Menlo Park, CA: AAAI Press.

Lavrenko, V., Lawrie, D., Ogilvie, P., and Schmill, M. 2003. Electronic Analyst of Stock Behavior. *Information Mining Seminar*. Amherst: Computer Science Department, University of Massachusetts. [Online]
Available from: http://ciir.cs.umass.edu/~lavrenko/aenalyst/index-old.html

Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D., and Allan J., 2000. Language Models for Financial News Recommendation. In *Proceedings of the 9th*

International Conference on Information and Knowledge Management (CIKM) (McLean, Virginia, November 06-11, 2000). New York: ACM Press, 2000, pp.389-396.

Lavrenko, V., Schmill, M., Lawire, D., Ogivie, P., Jensen, D., and Allan, J., 2000. Mining of Concurrent Text and Time Series. In *Proceedings of the Workshop on Text Mining at the Sixth International Conference on Knowledge Discovery and Data Mining, (Boston, MA, August 20-23, 2000).* New York: ACM Press, 2000, pp.37-44.

Law, M.H., Mario, Figueiredo, M.A.T., and Jain, A.K., 2002. *Feature Saliency in Unsupervised Learning*. Technical Report. East Lansing, Michigan: Department of Computer Science and Engineering, Michigan State University. [Online]
Available from: http://www.cse.msu.edu/~lawhiu/papers/TR02.ps.Z

Lee, K.S., and Kageura, K., 2006. Virtual Relevant Documents in Text Categorization with Support Vector Machines. *Information Processing and Management*. Article in Press, Now Available [Online]. Available from: http://www.sciencedirect.com

Lee, L.W., and Chen, S.M., 2006. New Methods for Text Categorization Based on a New Feature Selection Method and a New Similarity Measure Between Documents. In M. Ali and R. Dapoigny eds. *Advances in Applied Artificial Intelligence*, Proceedings of the *19th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (IEA/AEI), (Annecy, France, June 27-30, 2006), Lecture Notes in Computer Science*. Heidelberg, Berlin: Springer Verlag, 2006, vol.4031, pp.1280-1289.

Lewis, D.D. 1992. Feature Selection and Feature Extraction for Text Categorization. In *Proceedings of the Conference on Human Language Technology, Workshop on Speech and Natural Language (Harriman, New York, February 23-26, 1992).* Morristown, New Jersey (NJ): Association for Computational Linguistics, 1992, pp.212-217.

Lewis, D.D., and Ringuette, M., 1994. A Comparison of Two Learning Algorithms for Text Categorization. In *Proceedings of the 3rd Annual Symposium on Document Analysis and Information retrieval (SDAIR), (Las Vegas, Nevada, April 11-13, 1994).* pp.81-93.

Lin, J., and Gunopulos, D., 2003. Dimensionality Reduction by Random Projection and Latent Semantic Indexing. In D. Barbará and C. Kamath eds. *Text Mining Workshop* at *the 3rd SIAM International Conference on Data Mining, (San Francisco, May 1-3, 2003).*

Liu, H. and Motoda, H., 1998. *Feature Extraction, Construction and Selection: A Data Mining Perspective*. Boston, Massachusetts (MA): Kluwer Academic Publishers.

Liu, L., Kang, J., Yu, J., and Wang, Z., 2005. A Comparative Study on Unsupervised Feature Selection Methods for Text Clustering. In *Proceedings of the 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE), (Wuhan, China, Oct. 30-Nov. 01, 2005).* New York: IEEE, 2005, pp.597- 601.

Liu, T., Liu, S., Chen, Z. and Ma, W., 2003. An Evaluation on Feature Selection for Text Clustering. In T. Fawcett and N. Mishra eds. *Proceedings of the 20th International Conference on Machine Learning (ICML), (Washington, DC, August 21-24, 2003)*. Menlo Park, CA: AAAI Press, 2003, pp.488-495.

Loether, H.J., and McTavish, D.G., 1993. *Descriptive and Inferential Statistics: An Introduction*. 4th ed. Boston: Allyn and Bacon Inc.

Lowe, D., and Webb, A.R., 1991. Time Series Prediction by Adaptive Networks: a Dynamical Systems Perspective. In *Radar and Signal Processing, IEE Proceeding F*. 138(1), pp.17-24.

Malhi, A., and Gao, R.X., 2004. PCA-Based Feature Selection Scheme for Machine Defect Classification. *IEEE Transactions on Instrumentation and Measurement*, 53(6), pp. 1517-1525.

Malhotra, N.K., and Birks, D.F., 2002. *Marketing Research: An Applied Approach*. 2nd European ed. New Jersey (NJ): Financial Times/Prentice Hall Publishing.

Malkiel, B.G., 1996. *Random Walk Down Wall Street*. 6th ed. London: W.W. Norton Co.

Manomaisupat, P., and Abmad, K., 2005. Feature Selection for Text Categorization Using Self-organizing Map. In M. Zhao and Z, Shi eds. *Proceedings of the 2nd International Conference on Neural Networks and Brain (ICNN&B), (Beijing, China, October 13-15, 2005)*. IEEE Press, 2005, vol.3, pp.1875-1880.

Markellos, K., Markellou, P., Rigou, M., and Sirmakessis, S., 2003. Mining for Gems of Information. In S. Sirmakessis Ed. *Studies in Fuzziness and Soft Computing, Text Mining and its Applications: Results of the NEMIS Launch Conference on the 1st International Workshop on Text Mining and its Applications (Patras, Greece, April 5th, 2003)*. Berlin, Heidelberg: Springer-Verlag, 2004, Vol.138, pp.1-11.

Masuyama, T., and Nakagawa, H., 2002. Applying Cascaded Feature Selection to SVM Text Categorization. In A.M. Tjoa and R.R. Wagner eds. *Proceedings of the 13th International Workshop on Database and Expert Systems Applications, (Aix-en-Provence, France, Sep. 02-06, 2002)*. Washington, DC: IEEE Computer Society, 2002, pp.241-245.

McCallum, A.K., 1996. Bow: A Toolkit for Statistical Language Modeling, Text Retrieval, Classification, and Clustering. [Online].
Available from: http://www.cs.cmu.edu/~mccallum/bow

McCallum, A.K., and Nigam, K., 1998. A Comparison of Event Models for Naïve Bayes Text Classification. In ICML/AAAI *Workshop on Learning for Text Categorization (Wisconsin, July 26-27, 1998)* at *the 15th International Conference on Machine Learning*.

Merrill Lynch, Nov., 2000. e-Business Analytics: in Depth Report.

Mitchell, M.L., Mulherin, J.H., 1994. The Impact of Public Information on the Stock Market. *Journal of Finance*, 49(3), pp.923-950

Mitchell, T., 1996. *Machine Learning*. New York (NY): McGraw Hill.

Mittermayer, M.A., 2004. Forecasting Intraday Stock Price Trends with Text Mining Techniques. In *Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS) (Big Island, Hawaii, January 05-08, 2004)*. Washington, DC: IEEE Computer Society, 2004, 3(3), pp.30064.2.

Mladenic, D., 1998. Feature Subset Selection in Text-Learning. In C. Nedellec and C. Rouveirol eds. *Proceedings of the 10th European Conference on Machine Learning (ECML) (Chemnitz, Germany, April 21-23, 1998), Lecture Notes In Computer Science*. Heidelberg, Berlin: Springer-Verlag, 1998, vol.1398, pp.95-100.

Montanes, E., Combarro, E.F., Diaz, I., Ranilla, J., and Quevedo J.R., 2004. Words as Rules: Feature Selection in Text Categorization. In M. Bubak, G.D. van Albada, P.M.A. Sloot, and J. Dongarra eds. *Computational Sciences, Proceedings of the 4th International Conference on Computational Science, (Krakow, Poland, June 6-9, 2004), Lecture Notes in Computer Science*. Heidelberg, Berlin: Springer-Verlag, 2004, vol.3036, pp.666-669.

Montanes, E., Fernandez, J., Diaz, I., Combarro, E.F. and Ranilla, J., 2003. Measures of Rule Quality for Feature Selection in Text Categorization. In F. Pfenning, M.R. Berthold, H.J. Lenz, E. Bradley, R. Kruse, and C. Borgelt eds. *Advances in Intelligent Data Analysis V: Proceedings of the 5th International Symposium on Intelligent Data Analysis (IDA), (Berlin, Germany, August 28-30, 2003), Lecture Notes in Computer Science*. Heidelberg, Berlin: Springer-Verlag, 2003, vol.2810, pp.589-598.

Montgomery, D.C., and Runger, G.C., 1999. *Applied Statistics and Probability for Engineers*. 2nd ed. New York (NY): Wiley.

Morse, B.S., 2000. Lecture 18: Segmentation (Region Based). *Lecture Notes*. Hawaii: Brigham Young University. [Online]
Available from: homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/MORSE/region.pdf

Moyotl-Hernandez, E., and Jimenez-Salazar, H., 2005. Enhancement of DTP Feature Selection Method for Text Categorization. In A.F. Gelbukh ed. *Proceedings of 6th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing), (Mexico City, Mexico, February 13-19, 2005), Lecture Notes in Computer Science*. Heidelberg, Berlin: Springer-Verlag, 2005, vol.3406, pp.719-722.

Ng, A., and Fu, A.W., 2003. Mining Frequent Episodes for Relating Financial Events and Stock Trends. In K.Y. Whang, J. Jeon, K. Shim, and J. Srivastava eds. *Proceedings of the 7th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD, April 30 - May 2, 2003), Lecture Notes in Computer Science*. Heidelberg, Berlin: Springer-Verlag, 2003, Vol.2637, pp.27-39.

Ng, H.T., Goh, W.B., and Low, K.L., 1997. Feature Selection, Perception Learning, and a Usability Case Study for Text Categorization. In N.J. Belkin, A.D. Narasimhalu, P. Willett, and W. Hersh eds. *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*Philadelphia, Pennsylvania, July 27-31, 1997*). New York (NY): ACM Press, pp. 67-73.

Nigam, K., Lafferty, J., and McCallum, A., 1999. Using Maximum Entropy for Text Classification. In *Proceeding of the 16th International Joint Conference on Artificial Intelligence (IJCAI), Workshop on Machine Learning for Information Filtering (Stockholm, Sweden, July 31-August 6,* 1999), pp. 61-67.

Novovicova, J., and Malik, A., 2005. Information-Theoretic Feature Selection Algorithms for Text Classification. In *Proceedings of 2005 IEEE International Joint Conference on Neural Networks (IJCNN), (Montreal, Canada, July 31- August 04, 2005)*. New York (NY): IEEE, vol.5, pp.3272-3277.

Ogden, R.T., and Sugiura, N., 1994. Testing Change-points with Linear Trend. *Communications in Statistics B: Simulation and Computation*, 23(2), pp.287-322.

PaaB, G., Kindermann, J., and Leoold, E., 2003. Text Classification of News Articles with Support Vector Machines. In S. Sirmakessis Ed. *Studies in Fuzziness and Soft Computing, Text Mining and its Applications: Results of the NEMIS Launch Conference on the 1st International Workshop on Text Mining and its Applications (Patras, Greece, April 5th, 2003)*. Berlin, Heidelberg: Springer-Verlag, 2004, Vol.138, pp.53-64.

Papadimitriou, C.H., Raghavan, P., Tamaki, H., Vempala, S., 1998. *Latent Semantic Indexing: A Probabilistic Analysis*. In *Proceedings of the 17th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (Seattle, Washington, June 01-03, 1998)*. New York (NY): ACM Press, 1998, pp.159-168.

Parker, J., Sloan, T.M., and Yau, H., 1998. Data Mining. EPCC Technology Watch Report. Edinburgh Parallel Computing Center (EPCC), University of Edinburgh. [Online] Available from: http://www.epcc.ed.ac.uk/

Pavlidis, T., Horowitz, S.L., 1974. Segmentation of Plane Curves. *IEEE Transactions on Computers*, C-23(8), pp.860-870.

Permunetilleke, D., and Wong, R.K., 2002. Currency Exchange Rate Forecasting from News Headlines. In X. Zhou ed. *Proceedings of the 13h Australasian Conference on Database Technologies, Research and Practice in Information Technology Series, (Melbourne, Australia, January 28-February 2, 2002)*. Melbourne, Victoria: Australian Computer Society (ACS) Inc., 2002, Vol. 5, pp.131-139.

Phung, Y.C., 2005. Text Mining for Stock Movement Predictions: a Malaysian Perspective. In A. Zanasi, C.A. Brebbia and N.F.F. Ebecken eds. *Data Mining VI*,

Proceedings *of the 6th International Conference on Data Mining, Text Mining and Their Business Applications (Skiathos, Greece, May 25-27, 2004).* Southampton, Boston: WIT Press, 2005, vol.35, pp.103-111.

Ponte, J.M., and Croft, W.B., 1998. A Language Modeling Approach to Information Retrieval, In *Proceedings of the 21st Annual International Conference on Research and Development in Information Retrieval (Melbourne, Australia, August 24-28, 1998)*, New York (NY): ACM Press, 1998, pp.275-281.

Prabowo, R., and Thelwall, M., 2006. A Comparison of Feature Selection Methods for an Evolving RSS Feed Corpus. *International Journal of Information Processing and Management*, 42(6), pp.1491-1512.

Pring, M.J., 1991. T*echnical Analysis Explained*. New York (NY): McGraw-Hill.

Python Language Programming. 1990. [Online]. Available from: http://www.python.org

Quinlan, J.R., 1986. Induction of Decision Trees. *Machine Learning*, 1(1), pp.81-106.

R Project for Statistical Computing. 2003. Available form: http://www.r-project.org

Raghavan, P., 2002. Structure in Text: Extraction and Exploitation. In S. Amer-Yahia and L. Gravano eds. *Proceedings of the 7th International Workshop on the Web and Databases (WebDB): Collocated with ACM SIGMOD/PODS 2004 (Maison de la Chimie, Paris, France, June 17 - 18, 2004)*. New York (NY): ACM Press, 2004, Vol. 67. Keynote Talk available from: http://webdb2004.cs.columbia.edu/keynote.pdf

Rogati, M., and Yang, Y., 2002. High-Performing Feature Selection for Text Classification. In C. Nicholas, D.Grossman, K. Kalpakis, S. Qureshi, H. van Dissel, and L. Seligman eds. *Proceedings of the 11th International Conference on Information and Knowledge Management (CIKM), (McLean, Virginia, November 04-09, 2002)*. New York (NY): ACM Press, 2002, p.659-661.

Ruiz, M.E., and Srinivasan, P., 1999. Hierarchical Neural Networks for Text Categorization. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Berkeley, California, August 15-19, 1999)*. New York (NY): ACM Press, pp.281-282.

Salton, G., 1989. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Boston, Massachusetts: Addison-Wesley Longman Publishing Co., Inc.

Salton, G., and McGill, M.J., 1983. *An Introduction to Modern Information Retrieval*. New York (NY): McGraw-Hill.

Salton, G., and Yang, C.S., 1973. On the Specification of Term Values in Automatic Indexing. *Journal of Documentation*, 29(4), pp.351-372.

Schumaker, R.P., and Chen, H., 2006. Textual Analysis of Stock Market Prediction Using Financial News Articles. On the *12th Americas Conference on Information Systems (AMCIS, Acapulco, Mexico, August 4-6, 2006*).

Schutze, H., Hull, D.A., and Pedersen, J.O., 1995. Toward Optimal Feature Selection. In E.A. Fox, P. Ingwersen, and R. Fidel eds. *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Seattle, Washington, July 09-13, 1995)*. New York (NY): ACM Press, 1995, pp.229-237.

Sebastiani, F., 1999. A Tutorial on Automated Text Categorization. In A. Amandi and A. Zunino eds. *Proceedings of the 1st Argentinean Symposium on Artificial Intelligence (ASAI),* (*Buenos Aires, Argentina, September 08-09, 1999)*. pp. 7-35.

Sebastiani, F. 2002. Machine Learning in Automated Text Categorization. *ACM Computing Surveys (CSUR)*, 34(1), pp.1-47.

Seo, Y.W., Ankolekar, A., and Sycara, K., 2004. *Feature Selection for Extracting Semantically Rich Words*. Technical Report CMU-RI-TR-04-18. Pittsburgh, Pennsylvania: Robotics Institute, Carnegie Mellon University.

Shang, W., Huang, H., Zhu, H., Lin, Y., Qu, Y., and Wang, Z., 2006. A Novel Feature Selection Algorithm for Text Categorization. *Elsevier, Science Direct, Expert Systems with Applications*, 33(1), pp.1-5.

Shatkay , H., 1995. *Approximate Queries and Representations for Large Data Sequences*. Technical Report cs-95-03. Providence, Road Island (RI): Department of Computer Science, Brown University.

Shatkay, H., and Zdonik, S.B., 1996. Approximate Queries and Representations for Large Data Sequences. In S.Y. Su ed. *Proceedings of the Twelfth International Conference on Data Engineering (IDCE) (New Orleans, Louisiana, February 26-March 01, 1996)*. Washington, DC: IEEE Computer Society, 1996, pp.536-545.

Shaw, S.W., and deFigueiredo, R.J., 1990. Structural Processing of Waveforms as Trees. In *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(2), pp.328-338.

Siolas, G., and d'Alche-Buc, F., 2000. Support Vector Machines Based on a Semantic Kernel for Text Categorization. In S.I. Amari, C.L. Giles, M. Gori, and V. Piuri eds. *Proceedings of IEEE-INNS-ENNS International Joint Conference on Neural Networks,, Neural Computing: New Challenges and Perspectives for the New Millennium (Como, Italy, July 24-27, 2000)*. Washington: IEEE Computer Society, 2000, vol.5, pp.205-209.

Smith, L.I., 2002. A Tutorial on Principal Components Analysis. New Zealand: Department of Computer Science, University of Otago. [Online]
Available from: http://csnet.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

Song, F., Liu, S., and Yang, J., 2005. A Comparative Study on Text Representation Schemes in Text Categorization. *Pattern Analysis & Applications*, 8(1-2), pp.199-209.

Soucy, P., and Mineau, G.W., 2003. Feature Selection Strategies for Text Categorization. In Y. Xiang and B. Chaib-draa eds. *Advances in Artificial Intelligence, Proceedings of the 16th Conference of the Canadian Society for Computational Studies of Intelligence (Halifax, Canada, June 11-13, 2003), Lecture Notes in Computer Science*. Heidelberg, Berlin: Springer-Verlag, 2003, vol.2671, pp.505-509.

Sripada, S., Reiter, E., Hunter J., and Yu, J., 2002. Segmenting Time Series for Weather Forecasting. In X.A. Macintosh, R. Ellis, and F. Coenon eds. *Applications and Innovations in Intelligent Systems X, Proceedings of the 22nd SGAI International Conference on Knowledge Based Systems and Applied Artificial Intelligence (Cambridge, UK, December 10-12, 2002)*. New York (NY): Springer-Verlag, 2002, pp.193-206.

Steinbach, M., Karypis, G., and Kumar, V., 2000. A Comparison of Document Clustering Techniques. Poster in the *6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Workshop on Text Mining (Boston, MA, Aug. 20-23, 2000)*.

Sullivan, D., 2000. The Need for Text Mining in Business Intelligence. *DM Review Magazine*. December 2000 Issue. [Online]
Available from: http://www.dmreview.com/article_sub.cfm?articleId=2791

SVM Portal, 2005. *Optimum Separation Hyperplane*. [Online]
Available from: http://www.support-vector-machines.org/SVM_osh.html

Swan, R., and Allan, J., 1999. Extracting Significant Time Varying Features from Text. In S. Gauch and I.Y. Soong eds. *Proceedings of the 8th International Conference on Information and Knowledge Management (CIKM), (Kansas City, Missouri, November 02-06, 1999)*. New York (NY): ACM Press, 1999, pp.38-45.

Swan, R., and Allan, J., 2000. Automatic Generation of Overview Timelines. In E. Yannakoudakis, N.J. Belkin, M.K. Leong, and P. Ingwersen eds. *Proceedings of the 23rd Annual International SIGIR Conference on Research and Development in Information Retrieval, (Athens, Greece, July 24-28, 2000)*. New York: ACM Press, 1999, pp.49-56.

Tan, A. H., 1999. Text Mining: The State of Art and the Challenges. In *PAKDD Workshop on Knowledge Discovery from Advanced Databases (KDAD'99) in Conjunction with Third Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'9, Beijing, China, April 26-28, 1999)*. pp. 71-76.

Tang, B., Luo, X., Heywood, M.I., and Shepherd, M., 2004. *A Comparative Study of Dimension Reduction Techniques for Document Clustering.* Technical Report CS-2004-14. Nova Scotia, Canada: Faculty of Computer Science, Dalhousie University. [Online] Available from: http://www.cs.dal.ca/research/techreports/2004/CS-2004-14.pdf

Tay, F.E.H., Shen, L., and Cao, L., 2003. *Ordinary Shares, Exotic Methods: Financial Forecasting Using Data Mining Techniques.* River Edge, New Jersey (NJ): World Scientific Publishing Co., Inc.

Tehran Stock Exchange (TSE). 2005. *Introduction to Tehran Stock Exchange*. [Online]. Available from:http://www.tse.ir/qtp_27-04-2048/tse/Intro/intro.htm [cited January 2007].

Thomas, J.D., and Sycara, K., 2000. Integrating Genetic Algorithms and Text Learning for Financial Prediction. In A.A. Freitas, W. Hart, N. Krasnogor, and J. Smith eds. In *Data Mining with Evolutionary Algorithms, Proceedings of the Genetic and Evolutionary Computing Conference (GECCO) (Las Vegas, Nevada, July 8-12, 2000)*, pp.72-75.

Thomsett, M.C., 1998. *Mastering Fundamental Analysis*. Chicago: Dearborn Publishing.

Tokunaga, T., and Iwayama, M., 1994. *Text Categorization Based on Weighted Inverse Document Frequency*. Technical Report 94-TR0001. Tokyo, Japan: Department of Computer Science, Tokyo Institute of Technology.

Tzeras, K., and Hartman, S., 1993. Automatic Indexing Based on Bayesian Inference Networks. In R. Korfhage, E.M. Rasmussen, and P. Willett eds. *Proceedings of the 16th International ACM SIGIR Conference on Research and Development in Information Retrieval (Pennsylvania, June 27-July 01, 1993),* New York: ACM Press, 1993, pp.22-34.

Van Bunningen, A.H., 2004. *Augmented Trading - From News Articles to Stock Price Predictions Using Syntactic Analysis.* Master's Thesis. Enschede: University of Twente.

Van Rijsbergen, C.J., 1979. *Information Retrieval*. 2nd ed. London: Butterworths.

Vapnik, V.N., 1995. *The Nature of Statistical Learning Theory*. New York: Springer.

Vapnik, V.N., 1998. *Statistical Learning Theory*. New York (NY): Wiley-Inter Science.

Vempala, S., 1998. Random Projection: A New Approach to VLSI Layout. In *Proceedings of the 39th Annual Symposium on Foundations of Computer Science, (Palo Alto, CA, Nov. 08-11, 1998).* Washington: IEEE Computer Society, 1998, pp.389-395.

Vinay, V., Cox, I.J., Wood, K., and Milic-Frayling, N., 2005. A Comparison of Dimensionality Reduction Techniques for Text Retrieval. In *Proceedings of the 4th International Conference on Machine Learning and Applications (ICMLA), (Los Angles, California, Dec. 15-17, 2005).* Washington, DC: IEEE Computer Society, pp.293-298.

Wallis, F., Jin, H., Sista, S., and Schwartz, R., 1999. Topic Detection in Broadcast News. In *proceeding of the DARPA Broadcast News Workshop (Herndon, Virginia, February 28-March 3, 1999)*. [Online].
Available from: http://www.nist.gov/speech/publications/darpa99/html/tdt320/tdt320.htm

Wang, C., and Wang, X.S., 2000. Supporting Content-based Searches on Time Series via Approximation. In *Proceedings of the 12th International Conference on Scientific and Statistical Database Management (SSDBM), (Berlin, Germany, July 26-28, 2000)*. Washington, DC: IEEE Computer Society, 2000, pp.69-81.

Wang, Q., Guan, Y., Wang, X., and Xu, Z., 2006. A Novel Feature Selection Method Based on Category Information Analysis for Class Prejudging in Text Classification. *International Journal of Computer Science and Network Security,* 6(1A), pp.113-119

Wang, Y., and Wang, X.J., 2005. A New Approach to Feature selection in Text Classification. In *Proceedings of the 4th International Conference on Machine Learning and Cybernetics (Guangzhou, China, August 18-21, 2005)*. IEEE, vol.6, pp.3814-3819.

Wen, Y., 2001. *Text Mining Using HMM and PPM*. Unpublished Master's Thesis. Hamilton: University of Waikato.

Wen, Y., Witten, I.H., and Wang, D., 2003. Token Identification Using HMM and PPM Models. In T.D. Gedeon and L.C.C. Fung eds. *AI2003: Advances in Artificial Intelligence, Proceedings of the 16th Australian Conference on Artificial Intelligence (Perth, Australia, December 3-5, 2003), Lecture Notes in Computer Science*. Heidelberg, Berlin: Springer Verlag, 2003, vol.2903, pp.173-185.

White, H., 1988. Economic Prediction Using Neural Networks: The Case of IBM Daily Stock Returns. In *IEEE International Conference on Neural Networks (San Diego, California, July 24-27, 1988),* IEEE Press, 1988, Vol.2, pp.451-459.

Wiener, E., Pedersen, J.O., and Weigend, A.S., 1995. A Neural Network Approach to Topic Spotting. In *Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval (SDAIR),* (*Las Vegas, Nevada, April 24-26, 1995)*. pp.317-332.

Wikipedia, the Free Encyclopedia. 2001b. *Bi-gram and N-gram Definitions*. [Online]
Available from: http://en.wikipedia.org/wiki/bigram &http://en.wikipedia.org/wiki/ngram

Wikipedia, the Free Encyclopedia. 2001a. *Tokenization Definition*. [Online]
Available from: http://en.wikipedia.org/wiki/Tokenization [cited in November 2006]

Wilbur, J.W., and Sirotkin, K., 1992. The Automatic Identification of Stop Words. *Journal of Information Science,* 18(1), pp.45–55.

Witten, I.H., Paynter, G.W., Frank, E., Gutwin, C., and Nevill-Manning, C.G., 1999. KEA: Practical Automatic Keyphrase Extraction. Accepted Poster in *Proceedings of the*

4th International ACM Conference on Digital Libraries (Berkley, California, August 11-14, 1999). New York (NY): ACM Press, 1999, pp. 254-255.

Wu, X., 1993. Adaptive Split-and-Merge Segmentation Based on Piecewise Least Square Approximation. *IEEE Transactions on Pattern Analysis and Matching Intelligence, 1*5 (8), pp.808-815.

Wuthrich, B., Permunetilleke, D., Leung, S., Cho, V., Zhang, J., and Lam, W., 1998. Daily Stock Market Forecast from Textual Web Data. In *IEEE International Conference on Systems, Man, and Cybernetics (San Diego, California, October 11-14, 1998)*. IEEE Press, Vol.3, pp.2720-2725.

Wuthrich, B., 1997. Discovering Probabilistic Decision Rules. *International Journal of Intelligent Systems in Accounting Finance and Management*. New York (NY): John Wiley & Sons, Inc., 1997, 6(4), pp.269-277.

Wuthrich, B., 1995. Probabilistic Knowledge Bases. *IEEE Transactions* of *Knowledge and Data Engineering*. Piscataway, New Jersey (NJ): IEEE Educational Activities Department, 1995, 7(5), pp.691-698.

Wyse, N., Dubes, R., and Jain, A.K., 1980. A Critical Evaluation of Intrinsic Dimensionality Algorithms. In E. Gelsema and L. Kanal eds. *Pattern Recognition in Practice*. New York (NY): North-Holland Publishing Co., 1980, pp.415-425.

Yan, J., Liu, N., Zhang, B., Yan, S., Chen, Z., Cheng, Q., Fan, W., and Ma, W., 2005. OCFS: Optimal Orthogonal Centroid Feature Selection for Text Categorization. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Salvador, Brazil, August 15-19, 2005)*. New York (NY): ACM Press, 2005, pp.122-129.

Yang, H.S., and Lee, S.U., 1997. Split-and-Merge Segmentation Employing Thresholding Technique. In *Proceedings of the 1997 International Conference on Image Processing (ICIP), (Washington, DC, October 26-29, 1997)*. Washington, DC: IEEE Computer Society, 1997, vol. 1, pp.239-242.

Yang, Y., 1994. Expert Network: Effective and Efficient Learning from Human Decisions in Text Categorization and Retrieval. In W.B. Croft and C.J. van Rijsbergen eds. *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Dublin, Ireland, July 03-06, 1994)*. New York (NY): Springer-Verlag, 1994, pp.13-22

Yang, Y., 1995. Noise Reduction in a Statistical Approach to Text Categorization. In E.A. Fox, P. Ingwersen, and R. Fidel eds. *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Seattle, Washington, July 09-13, 1995)*. New York (NY): ACM Press, 1995, pp.256-263.

Yang, Y., 1999. An Evaluation of Statistical Approaches to Text Categorization. *Information Retrieval*, 1(1-2), pp.69-90.

Yang, Y., and Chute, C.G., 1994. An Example-based Mapping Method for Text Categorization and Retrieval. *ACM Transactions on Information Systems (TOIS): Special Issue on Text Categorization*, 12(3), pp.252-277.

Yang, Y., Liu, X., 1999. A Re-Examination of Text Categorization Methods. In Proceedings *of the 22nd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (Berkley, California, August 15-19, 1999)*. New York (NY): ACM Press, 1999, pp.42-49.

Yang, Y., and Pedersen, J.O., 1997. A Comparative Study on Feature Selection in Text Categorization. In D.H. Fisher ed. *Proceedings of the 14th International Conference on Machine Learning (ICML), (Nashville, Tennessee, July 08-12, 1997)*. San Francisco, California: Morgan Kaufmann Publishers Inc., 1997, pp.412-420.

Yang, Y., and Wilbur, J., 1996. Using Corpus Statistics to Remove Redundant Words in Text Categorization. *Journal of the American Society for Information Science*, 47(5), pp.357-369.

Yilmazel, O., Symonenko, S., Balasubramanian, N., and Liddy, E.D., 2005. Improved Document Representation for Classification Tasks for the Intelligence Community. In *Technical Report SS-05-01. Proceedings of AAAI 2005 Spring Symposium on AI Technologies for Homeland Security (Stanford, California, March 21-23, 2005)*. Menlo Park, California: AAAI Press, 2005, pp.76-82.

Yu, J.X., Ng, M.K., and Huang, J.Z., 2001. Patterns Discovery Based on Time-Series Decomposition. In D.W. Cheung, G.J. Williams, and Q. Li, eds. *Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Lecture Notes in Computer Science (Hong Kong, China, April 16-18, 2001)*. London: Springer-Verlag, 2001, vol.2035, pp.336-347.

Zhang, T. and J. Oles, F. 2001. Text Categorization Based on Regularized Linear Classification Methods. *Information Retrieval*, 4(1), pp.5-31.

Zheng, Z., Srihari, R.K, 2003. Optimally Combining Positive and Negative Features for Text Categorization. *Workshop on Learning from Imbalanced Datasets II, Proceedings of 20th International Conference on Machine Learning, (Washington, Aug. 21-24, 2003)*.

Zheng, Z., Srihari, R.K, and Srihari, S.N., 2003. A Feature Selection Framework for Text Filtering. In *Proceedings of the 3rd IEEE International Conference on Data mining (ICDM), (Melbourne, Florida, November 19-22, 2003)*. Washington, DC: IEEE Computer Society, 2003, p. 705-708.

Zheng, Z., Wu, X., and Srihari, R., 2004. Feature Selection for Text Categorization on Imbalanced Data. *ACM SIGKDD Explorations Newsletter*, 6(1), pp.80-89.
Zhou, Z.H., 2003. Three Perspectives of Data Mining. *Journal of Artificial Intelligence,* 143(1), pp.139-146.

Zorn, P., Emanoil, M., Marshall, L., and Panek, M., 1999. Mining Meets the Web. *Online,* 23(5), pp.17-28.

Appendix 1: The 15 Selected Online News Sources

Online Time-Stamped News Sources	Website Link	Number of News
خبرگزاري كار ايران - ايلنا ILNA - Iranian Labor News Agency	http://www.ilna.ir	147
خبرگزاري دانشگاه آزاد اسلامي - آنا ANA - Azad News Agency	http://www.ana.ir	125
خبرگزاري بورس تهران - بورس نيوز Bourse News	http://www.boursenews.ir	32
خبرگزاري دانشجويان ايران - ايسنا ISNA - Iranian Students News Agency	http://www.isna.ir	182
باشگاه خبرنگاران دانشجوي ايران ISCA - Iran Student Correspondents Association	http://www.iscanews.ir	61
خبرگزاری فارس FARS - Fars News Agency	http://www.farsnews.com	505
خبرگزاری مهر MEHR - Mehr News Agency	http://www.mehrnews.com	160
واحد مركزي خبر IRIB - Islamic Republic of Iran Broadcasting	http://www.irib.ir	3
خبرگزاري آريا Arya News Agency	http://www.aryanews.com	4
شبكه اطلاع رساني نفت و انرژي - شانا SHANA – Petroenergy Information Network	http://www.shana.ir	107
خبرگزاری اقتصادی ايران Economic News Agency	http://www.econews.ir	7
شريف نيوز -پايگاه تحليلی خبری دانشجويان ايران Sharif News	http://www.sharifnews.ir	54
آفتاب نيوز Aftab News	http://www.aftabnews.ir	29
سايت خبری بازتاب Baztab Professional News Site	http://www.baztab.ir	11
روزنامه آسيا Newspaper Asia	http://www.asianews.ir	96

Appendix 2: 1523 News Collected from 15 News Sources

Source : ILNA - Iranian Labor News Agency				
IL8412231204	IL8408211408	IL8404041652	IL8402191223	IL8311121750
IL8412201542	IL8406141620	IL8404041450	IL8402171609	IL8311071640
IL8412161506	IL8406051517	IL8404011420	IL8402061019	IL8311061342
IL8412141041	IL8405181537	IL8403291324	IL8401271503	IL8311061629
IL8412121116	IL8405181534	IL8403231156	IL8401141650	IL8311051605
IL8410060948	IL8405061249	IL8403221627	IL8401061233	IL8311051520
IL8410051325	IL8405050858	IL8403171013	IL8312201409	IL8311041433
IL8409221424	IL8404222033	IL8403161302	IL8312121735	IL8311031514
IL8409221231	IL8404221604	IL8403091016	IL8312061452	IL8311031450
IL8409201001	IL8404141532	IL8403081550	IL8311281506	IL8310280923
IL8409191536	IL8404141514	IL8403021243	IL8311281431	IL8310261255
IL8409181028	IL8404140928	IL8402312304	IL8311281442	IL8310211519
IL8409131311	IL8404091327	IL8402312306	IL8311251418	IL8310091355
IL8409051019	IL8404071431	IL8402281647	IL8311131656	IL8309291312
IL8306271311	IL8404061603	IL8402241649	IL8311121647	IL8309291351
IL8306240919	IL8306021451	IL8305251546	IL8305241110	IL8305211103
IL8306231136	IL8305271109	IL8305251158	IL8305221356	IL8305032025
IL8305031230	IL8304211450	IL8304081543	IL8303272020	IL8303161719
IL8305031056	IL8304212029	IL8304071314	IL8303231808	IL8303101728
IL8304271412	IL8304171511	IL8304051746	IL8303201744	IL8303071526
IL8304211424	IL8304161454	IL8304021111	IL8303201955	IL8303061726
IL8302231236	IL8302211511	IL8302201236	IL8302161540	IL8302131417
IL8302221255	IL8302211711	IL8302181131	IL8302131620	IL8302121445
IL8302080952	IL8302061316	IL8302021450	IL8301261110	IL8301261824
IL8301221226	IL8301061417	IL8301061159	IL8308171228	IL8303061721
IL8309292154	IL8309131426	IL8309031648	IL8308111523	IL8303021459
IL8309241040	IL8309111422	IL8309011901	IL8308051538	IL8302291359
IL8309241444	IL8309091915	IL8308291507	IL8307261110	IL8302251426
IL8309231137	IL8309081504	IL8308271136	IL8307141033	IL8302121657
IL8302081830	IL8301251507			
Source : Bourse News				
BO8412141308	BO8408110134	BO8407010334	BO8410181037	BO8405171856
BO8411271520	BO8408100550	BO8406280904	BO8409281517	BO8405110728
BO8411261534	BO8408031520	BO8406241833	BO8409271327	BO8404230035
BO8411031423	BO8407300131	BO8406241600	BO8409261300	BO8404171310
BO8410191145	BO8407270632	BO8406240807	BO8405271601	BO8405310935
BO8407030946	BO8407190536	BO8406202328	BO8406181754	BO8406122053
BO8407030339	BO8407090537			
Source : IRIB – Islamic Republic of Iran Broadcasting				
IR8410240815	IR8409301027	IR8411091644		

Source : ISNA - Iranian Students News Agency				
IS8412281545	IS8412031134	IS8410261345	IS8410051539	IS8409151208
IS8412221016	IS8412011512	IS8410241425	IS8410021245	IS8409011220
IS8412211351	IS8412011500	IS8410251246	IS8409301802	IS8409011003
IS8412171606	IS8411261329	IS8411041204	IS8409281506	IS8408301135
IS8412161720	IS8411261122	IS8410201422	IS8409231250	IS8408281522
IS8412121334	IS8411171332	IS8410191621	IS8409221212	IS8408271135
IS8412091153	IS8411151211	IS8410191648	IS8409211749	IS8408251607
IS8412061020	IS8411091701	IS8410172113	IS8409211654	IS8408251422
IS8412051329	IS8411081817	IS8410161038	IS8409201451	IS8408241152
IS8412031402	IS8411031732	IS8410161142	IS8409201109	IS8408181401
IS8407091326	IS8405251224	IS8404281243	IS8409151251	IS8404051042
IS8407051519	IS8405261228	IS8404281300	IS8402261347	IS8403171357
IS8407021602	IS8405211224	IS8404271125	IS8402261011	IS8403071256
IS8406311723	IS8405201241	IS8404261502	IS8402181331	IS8403062020
IS8406231446	IS8405161314	IS8404241120	IS8402141530	IS8403061758
IS8406201439	IS8405041356	IS8404241502	IS8402061315	IS8403061040
IS8406151109	IS8405031606	IS8404191244	IS8402031603	IS8403041036
IS8406111257	IS8405021322	IS8404101312	IS8402031501	IS8402301319
IS8405311410	IS8404291045	IS8404061104	IS8401301536	IS8402261635
IS8310291442	IS8310111248	IS8404051412	IS8401221133	IS8402261434
IS8310271631	IS8310130957	IS8309211559	IS8401201155	IS8308291418
IS8311041128	IS8310051601	IS8309151255	IS8309130227	IS8308261301
IS8310141728	IS8309301157	IS8309130057	IS8309011315	IS8308111402
IS8308062022	IS8307221801	IS8306291530	IS8306201431	IS8306101046
IS8308051558	IS8307171106	IS8306271712	IS8306181604	IS8306151110
IS8308011218	IS8307251211	IS8306201259	IS8306161041	IS8306060903
IS8305201247	IS8304311615	IS8304171233	IS8303272116	IS8303061517
IS8305090836	IS8304231008	IS8304131659	IS8303171654	IS8302301406
IS8305041109	IS8304271152	IS8304091643	IS8303081153	IS8302231108
IS8304201238	IS8304221034	IS8304021448	IS8303061659	IS8302261132
IS8301151207	IS8407091024	IS8407171132	IS8312091316	IS8311261447
IS8408231048	IS8407281447	IS8407121029	IS8312161221	IS8311131345
IS8408151040	IS8407261553	IS8401141936	IS8312051629	IS8311061814
IS8408081438	IS8407261353	IS8312251510	IS8312041241	IS8310291501
IS8408021421	IS8407191246	IS8312191411	IS8311271149	IS8308081201
IS8306011650	IS8306011235	IS8305281222	IS8302211758	IS8302120949
IS8302081112	IS8301181804			
Source : BAZTAB – Baztab Professional News Site				
BA8406311618	BA8401141516	BA8309171133	BA8408261020	BA8404131457
BA8406231802	BA8406231236	BA8403151548	BA8309301416	BA8403171523
BA8312220949				

Source : MEHR - Mehr News Agency				
ME8409281101	ME8405161525	ME8402141248	ME8410051547	ME8405211056
ME8304081452	ME8308121608	ME8402171230	ME8411021614	ME8311261451
ME8403281201	ME8411181021	ME8311031623	ME8409201316	ME8302231405
ME8404071717	ME8411091641	ME8404291010	ME8303191319	ME8302141317
ME8405161522	ME8410101625	ME8403291258	ME8303221021	ME8301091436
ME8303290930	ME8408081357	ME8403181602	ME8303301802	ME8303311326
ME8305211040	ME8308061555	ME8301151544	ME8302121338	ME8310011208
ME8305201203	ME8305231117	ME8410261154	ME8310220952	ME8310191530
ME8307201753	ME8302201631	ME8410291318	ME8406221513	ME8311061615
ME8407281235	ME8404311155	ME8301061346	ME8412161437	ME8401061345
ME8410131641	ME8311271254	ME8407021004	ME8410011707	ME8405191058
ME8308161448	ME8312031434	ME8402061158	ME8409261148	ME8405051009
ME8403231243	ME8311171952	ME8303301424	ME8402261109	ME8411261347
ME8407011145	ME8312031551	ME8304091206	ME8403041344	ME8310281747
ME8302191535	ME8310261611	ME8301181953	ME8401261119	ME8311041220
ME8302041447	ME8402110011	ME8301291318	ME8401281536	ME8405251650
ME8310151542	ME8402021421	ME8309101609	ME8311121505	ME8405311242
ME8309021449	ME8402061039	ME8309061111	ME8410181540	ME8403171246
ME8410031251	ME8406251000	ME8309241120	ME8409231313	ME8303301342
ME8410111516	ME8406250955	ME8309211459	ME8409251108	ME8309211521
ME8409151447	ME8406250950	ME8309231436	ME8409281505	ME8309081311
ME8411211240	ME8309091353	ME8402171055	ME8409291629	ME8308041343
ME8411261505	ME8308111454	ME8402251422	ME8411101415	ME8310011458
ME8409141240	ME8308051414	ME8402251519	ME8411151330	ME8404261036
ME8406181056	ME8407251334	ME8302031445	ME8403211210	ME8406231503
ME8406181055	ME8410051413	ME8406261043	ME8308061535	ME8412011517
ME8307251457	ME8410011712	ME8406041109	ME8308261607	ME8412141214
ME8310091915	ME8409191604	ME8404141358	ME8304242006	ME8409271154
ME8311061328	ME8302201421	ME8405121128	ME8301261505	ME8408151330
ME8401271149	ME8309231258	ME8308021358	ME8302191443	ME8408151537
ME8406231459	ME8307221714	ME8310151434	ME8410201206	ME8409181039
ME8310231457	ME8311251631	ME8310161143	ME8408281109	ME8311121221
Source : AFTAB - Aftab News Agency				
AF8408051425	AF8410291600	AF8409271025	AF8409041626	AF8404261802
AF8409171605	AF8411040817	AF8406110105	AF8412182231	AF8405011642
AF8404281227	AF8407092210	AF8405191134	AF8409051546	AF8405221819
AF8405171334	AF8412020938	AF8408251443	AF8408251437	AF8410302004
AF8409251301	AF8409222243	AF8408251435	AF8403161632	AF8411171840
AF8406110004	AF8407292237	AF8409011233	AF8412020859	
Source : ARYA - Arya News Agency				
AR8406191429	AR8406211629	AR8411171109	AR8411161039	

Source : ANA – Azad News Agency				
AN8412270823	AN8410271514	AN8405241113	AN8404051440	AN8403021343
AN8412241403	AN8410261505	AN8405191750	AN8404051054	AN8402041231
AN8412131102	AN8410241656	AN8405091300	AN8404011045	AN8312030726
AN8412071200	AN8410131546	AN8405081229	AN8403260759	AN8312030720
AN8412011347	AN8410121555	AN8405050922	AN8403231648	AN8311131555
AN8411151459	AN8410111627	AN8405011138	AN8403231647	AN8311131554
AN8411121606	AN8410061437	AN8404301302	AN8403211344	AN8311131210
AN8411101456	AN8410031505	AN8404301301	AN8403211343	AN8311121557
AN8411051524	AN8409291616	AN8404251316	AN8403171202	AN8311111554
AN8411031727	AN8409291052	AN8404211123	AN8302281032	AN8311071042
AN8411030856	AN8409261616	AN8404141757	AN8403161335	AN8311061824
AN8407301550	AN8409221501	AN8404121359	AN8403101207	AN8311061118
AN8407251651	AN8409201424	AN8404101013	AN8403081041	AN8311051038
AN8407241228	AN8409151551	AN8404061437	AN8304170903	AN8311041158
AN8407191159	AN8409151550	AN8306101721	AN8304170901	AN8311041157
AN8407181244	AN8409051532	AN8305311530	AN8304091533	AN8310251230
AN8407111545	AN8409021604	AN8305291030	AN8304091530	AN8305211058
AN8407021634	AN8408261239	AN8305281445	AN8304031456	AN8305051154
AN8405241114	AN8408081105	AN8305221528	AN8304170909	AN8304221646
AN8304021424	AN8304181116	AN8303061543	AN8303091556	AN8303231652
AN8304011001	AN8304171532	AN8302281348	AN8303091551	AN8303201514
AN8303231801	AN8309081012	AN8309011639	AN8307261623	AN8306281440
AN8310091526	AN8309051128	AN8308191657	AN8307231111	AN8306211508
AN8309131400	AN8309021657	AN8308071057	AN8306291635	AN8306181058
AN8309091540	AN8306141358	AN8304211513	AN8304211505	AN8304191219
Source : ISCA - Iran Student Correspondants Association				
IC8412261404	IC8411261444	IC8411031332	IC8408271339	IC8407171527
IC8412051421	IC8411261358	IC8410191715	IC8408181526	IC8407081531
IC8412031614	IC8411261357	IC8410191422	IC8408271157	IC8407011307
IC8412011331	IC8411251135	IC8410131720	IC8408181323	IC8406011620
IC8412011317	IC8411251105	IC8409291706	IC8408181538	IC8405161636
IC8412011316	IC8411231107	IC8409280919	IC8408021612	IC8404141721
IC8412011234	IC8411231049	IC8409201658	IC8407291118	IC8404121747
IC8411291419	IC8411221258	IC8409181125	IC8407261534	IC8403311101
IC8401262214	IC8411161216	IC8409151128	IC8407251146	IC8403291521
IC8411261501	IC8411111452	IC8409120908	IC8407191254	IC8403231749
IC8403231634	IC8403211722	IC8402311823	IC8402191612	IC8402141849
IC8403231434	IC8403211650	IC8402261718	IC8402191454	IC8402042105
IC8402042102				
Source : ECONEWS - Economic News Agency				
EC8412271430	EC8412201905	EC8412171908	EC8412162055	EC8412162029
EC8411300853	EC8412141519			

Source : SHANA - Petroenergy Information Network				
SH8412091350	SH8410111535	SH8409021902	SH8407301059	SH8406191030
SH8412011446	SH8408261524	SH8408291333	SH8407251328	SH8406121727
SH8411181044	SH8410101556	SH8408301331	SH8407131208	SH8406121710
SH8411051600	SH8410091214	SH8408251355	SH8407031259	SH8406091926
SH8411031551	SH8410091140	SH8408251331	SH8407021500	SH8406091916
SH8410301508	SH8410071046	SH8408231514	SH8407021253	SH8406091844
SH8410301456	SH8410051804	SH8408141635	SH8406301144	SH8406091341
SH8410261205	SH8410051716	SH8408120954	SH8406231725	SH8406011615
SH8410241559	SH8409271735	SH8407301118	SH8406211524	SH8405311833
SH8405311155	SH8404041824	SH8402281211	SH8312171611	SH8310281318
SH8405291913	SH8404021238	SH8402241615	SH8312061859	SH8310261143
SH8405191239	SH8403301629	SH8402141414	SH8312041348	SH8310191721
SH8405191234	SH8403291537	SH8402121312	SH8312041345	SH8310161558
SH8405171058	SH8403231203	SH8401281220	SH8312031328	SH8310161553
SH8405151253	SH8403231144	SH8401251445	SH8312031238	SH8310051528
SH8405111336	SH8403221334	SH8401211133	SH8311281654	SH8310011624
SH8405251648	SH8403171343	SH8401150939	SH8312021411	SH8309231712
SH8404121249	SH8403101550	SH8312261323	SH8311051515	SH8309221410
SH8403091540	SH8403091801	SH8312221624	SH8310291548	SH8309211612
SH8309151156	SH8309101716	SH8309071836	SH8309041648	SH8308131907
SH8309141256	SH8309091533	SH8309041701	SH8309031413	SH8308041335
SH8308301700	SH8308161253			
Source : ASIA Online Newspaper				
AS8410071537	AS8409211530	AS8409051654	AS8408290846	AS8408280846
AS8410051006	AS8409191011	AS8409071813	AS8409040759	AS8408240745
AS8410050859	AS8409210811	AS8409131710	AS8409041659	AS8408230855
AS8409230825	AS8409200845	AS8409040807	AS8408291535	AS8408201046
AS8409280708	AS8409140741	AS8409010747	AS8408271814	AS8408231703
AS8409180852	AS8409091025	AS8409011642	AS8408291631	AS8408080813
AS8408080821	AS8408061125	AS8407190915	AS8407091104	AS8407120955
AS8408061116	AS8407250911	AS8407111100	AS8407080935	AS8406260757
AS8408101126	AS8407261013	AS8407100854	AS8407061205	AS8406230947
AS8406260829	AS8405261503	AS8405071458	AS8405011655	AS8404121516
AS8407011029	AS8405291524	AS8405121451	AS8404261520	AS8404081443
AS8406221055	AS8405221556	AS8405081513	AS8404241325	AS8404061454
AS8406260758	AS8405161746	AS8405021555	AS8404201439	AS8404071441
AS8407041022	AS8405151706	AS8404311456	AS8404141453	AS8404061504
AS8406180826	AS8405151718	AS8404281514	AS8404141438	AS8404051529
AS8406180828	AS8405081450	AS8404271524	AS8404141443	AS8404051530
AS8405291506	AS8403241759	AS8403231421	AS8404121356	AS8404011445
AS8403221425	AS8403231358	AS8403171457	AS8403161527	AS8403181523
AS8402301507	AS8403021540	AS8402311425	AS8402041552	AS8402021401
AS8312082038				

Source : FARS – Fars News Agency				
FA8412241802	FA8409261142	FA8406121529	FA8402231809	FA8309091356
FA8412221816	FA8409241930	FA8406101827	FA8402191516	FA8309071759
FA8412212009	FA8409230829	FA8406101804	FA8402171300	FA8309011337
FA8412200947	FA8409201155	FA8406091818	FA8402031551	FA8308191825
FA8412191554	FA8409142147	FA8405311357	FA8401271128	FA8308191658
FA8412191445	FA8409141100	FA8405291713	FA8401201116	FA8308151035
FA8412171609	FA8409130043	FA8405291522	FA8401071245	FA8308131124
FA8412161600	FA8409111421	FA8405241832	FA8401061848	FA8308121008
FA8412071626	FA8409052203	FA8405171143	FA8312261440	FA8308112250
FA8412011631	FA8409040846	FA8405161608	FA8312251213	FA8308061127
FA8411301824	FA8409021725	FA8405151223	FA8312251208	FA8308041515
FA8411301820	FA8408281040	FA8405151033	FA8312241556	FA8308041409
FA8411241743	FA8408241758	FA8405111511	FA8312231835	FA8308021622
FA8411241530	FA8408111820	FA8405041205	FA8312221350	FA8307291834
FA8411141438	FA8408091332	FA8404311100	FA8312181445	FA8307221942
FA8411091640	FA8408060945	FA8404301315	FA8312151714	FA8307182254
FA8411051532	FA8407301046	FA8404271912	FA8312121429	FA8307120832
FA8411031731	FA8407301023	FA8404271010	FA8312111636	FA8306230013
FA8411031024	FA8407281512	FA8404231920	FA8312071038	FA8306181634
FA8410241804	FA8407241528	FA8404101412	FA8312061808	FA8306100711
FA8410161705	FA8407191023	FA8404101120	FA8312051205	FA8306011455
FA8410161616	FA8407091024	FA8404091820	FA8312041543	FA8306011148
FA8410121102	FA8407081028	FA8404081330	FA8312031453	FA8306312217
FA8410101832	FA8407051143	FA8404051808	FA8312031330	FA8305311034
FA8410061255	FA8406311123	FA8404051711	FA8312031054	FA8305281233
FA8410061200	FA8406301219	FA8404041229	FA8311281319	FA8305260940
FA8410051315	FA8406271146	FA8404011312	FA8311271151	FA8305251605
FA8410031718	FA8406221504	FA8403301805	FA8311241510	FA8305231131
FA8410021148	FA8406221300	FA8403301516	FA8311131641	FA8305221301
FA8410011718	FA8406141637	FA8403291537	FA8311071457	FA8305201525
FA8409291633	FA8406131524	FA8403291530	FA8311061435	FA8305051640
FA8409282148	FA8406131452	FA8403231244	FA8311051420	FA8305032119
FA8409281248	FA8406121604	FA8403221545	FA8311041351	FA8304301854
FA8409261701	FA8309291655	FA8403171206	FA8311031614	FA8304290145
FA8310121322	FA8309241703	FA8403161533	FA8310261155	FA8304271623
FA8310121250	FA8309241116	FA8403041609	FA8310251233	FA8304231908
FA8310081738	FA8309231247	FA8403031649	FA8310241452	FA8304231622
FA8310081733	FA8309221826	FA8403021602	FA8310231736	FA8304231502
FA8310081721	FA8309220947	FA8402311300	FA8310221655	FA8304211748
FA8310051526	FA8309220940	FA8402281314	FA8310221627	FA8304211656
FA8310031758	FA8309220938	FA8402251523	FA8310151708	FA8304191059
FA8309301524	FA8309211831	FA8402241602	FA8310141625	FA8304181633
FA8309291902	FA8309161748	FA8309241717	FA8312081220	FA8304171902

Source : FARS – Fars News Agency				
FA8304151840	FA8309141146	FA8309240840	FA8312051533	FA8411151451
FA8304151829	FA8309101832	FA8309221853	FA8312051346	FA8411121600
FA8304091244	FA8302291357	FA8309220942	FA8312031349	FA8411101440
FA8304090932	FA8302280925	FA8309211924	FA8311271506	FA8411100001
FA8304081648	FA8302261202	FA8309211755	FA8311270923	FA8411041849
FA8304061153	FA8302221354	FA8309161506	FA8311241559	FA8411031616
FA8304061139	FA8302211753	FA8309161030	FA8311241309	FA8410141639
FA8304042057	FA8302211349	FA8309141139	FA8311241259	FA8410131721
FA8304020934	FA8302201836	FA8309131324	FA8311201648	FA8410131253
FA8304010944	FA8302181554	FA8309091659	FA8311132012	FA8410091220
FA8303311535	FA8302131635	FA8309082048	FA8311111712	FA8410061112
FA8303301607	FA8302121907	FA8309071801	FA8311061109	FA8410052250
FA8303291254	FA8302121900	FA8309071426	FA8311041348	FA8410052228
FA8303291157	FA8302091704	FA8309011901	FA8311031606	FA8410051521
FA8303250907	FA8302082312	FA8309011814	FA8310291804	FA8410051510
FA8303211728	FA8301161055	FA8308301806	FA8310281631	FA8410051419
FA8303201623	FA8301151255	FA8308291125	FA8310281401	FA8410011720
FA8303201031	FA8312261455	FA8308271401	FA8310271549	FA8409291646
FA8303172205	FA8312181623	FA8308241057	FA8310231458	FA8409281636
FA8303161530	FA8312181455	FA8308191323	FA8310202126	FA8409271601
FA8303091254	FA8312161551	FA8308191257	FA8310161433	FA8409251032
FA8303041245	FA8312151606	FA8308171238	FA8310141533	FA8409130936
FA8302301834	FA8306291805	FA8308161450	FA8310091745	FA8409091917
FA8302291730	FA8306281441	FA8308061929	FA8310071510	FA8409052156
FA8304060915	FA8306271238	FA8308051511	FA8310071143	FA8408291955
FA8304021832	FA8306251903	FA8308031459	FA8310061559	FA8408281924
FA8304011323	FA8306211702	FA8308021417	FA8310061557	FA8408281700
FA8303311550	FA8306191849	FA8307071557	FA8310021521	FA8408281658
FA8303311419	FA8306171755	FA8307071033	FA8310021455	FA8408251350
FA8303301026	FA8306171420	FA8306301605	FA8310011655	FA8408251227
FA8303281251	FA8306161607	FA8301061541	FA8309301441	FA8408251208
FA8303251559	FA8306071020	FA8301051925	FA8309291758	FA8408241656
FA8303201615	FA8306051419	FA8412241236	FA8309281337	FA8408191521
FA8303081040	FA8305281230	FA8412161545	FA8406211217	FA8408121645
FA8303061623	FA8305281025	FA8412161503	FA8406211000	FA8408020933
FA8303021810	FA8305201601	FA8412091300	FA8406201557	FA8407271351
FA8302272038	FA8305181621	FA8412032242	FA8406181301	FA8407261103
FA8302261904	FA8305061426	FA8412011549	FA8406151707	FA8407261059
FA8302261221	FA8305041546	FA8411271500	FA8406091826	FA8407251824
FA8302221709	FA8304221740	FA8411261921	FA8406091822	FA8407121156
FA8302220935	FA8304212200	FA8411182245	FA8406051518	FA8407121135
FA8302131920	FA8304161826	FA8411180930	FA8406011654	FA8407081025
FA8302061855	FA8304130922	FA8406241215	FA8406011604	FA8406301208
FA8301261922	FA8304100834	FA8406231506	FA8405311247	FA8406291148

Source : FARS – Fars News Agency				
FA8301151251	FA8304071150	FA8406221459	FA8405311133	FA8406241225
FA8405251532	FA8404232010	FA8403111132	FA8405311035	FA8401311731
FA8405231307	FA8404151312	FA8403101433	FA8402211319	FA8401311312
FA8405221227	FA8404141026	FA8403091609	FA8402181320	FA8401311125
FA8405201549	FA8404121043	FA8403071421	FA8402171013	FA8401291253
FA8405191320	FA8404081329	FA8403071144	FA8402142233	FA8401251554
FA8405181522	FA8404071627	FA8403061107	FA8402111636	FA8401211215
FA8405121251	FA8404021834	FA8403051235	FA8402111631	FA8401211138
FA8405081223	FA8404011543	FA8403021551	FA8402101155	FA8401172028
FA8405041520	FA8403301500	FA8403021258	FA8402060005	FA8401141749
FA8405011346	FA8403231257	FA8403011520	FA8403171210	FA8402021252
FA8404281345	FA8403221459	FA8403011116	FA8402041732	FA8402041404
FA8404251100	FA8403210943	FA8402311908	FA8402041550	FA8402031319
FA8403181441	FA8402260950	FA8402261340	FA8402211716	FA8402261113
Source : SHARIF - Sharif News				
SA8310102033	SA8402021207	SA8405011236	SA8408241358	SA8410121136
SA8310141825	SA8402070510	SA8405071505	SA8408291151	SA8410131508
SA8310151033	SA8402071901	SA8405081140	SA8409161014	SA8410180802
SA8310291656	SA8402241718	SA8405192057	SA8409201505	SA8410261252
SA8312091649	SA8402251331	SA8405251523	SA8409271617	SA8411101144
SA8312150957	SA8403030527	SA8405302354	SA8409281633	SA8411151702
SA8312181741	SA8404121530	SA8406221649	SA8409291229	SA8412061423
SA8312151158	SA8404221627	SA8406252213	SA8410012215	SA8412141451
SA8401151048	SA8404250233	SA8406252214	SA8410031605	SA8412161831
SA8401241439	SA8404271325	SA8406311459	SA8410061318	SA8412290251
SA8401261147	SA8404290358	SA8408151117	SA8410101641	

Appendix 3: News and Trend Alignment Results

#	Segments Beginning & Ending Time		Trend /Doc		#	Segments Beginning & Ending Time		Trend /Doc	
1	8301081043	8301101229	1	1	43	8304011155	8304020901	1	1
2	8301151112	8301161138	0	5	44	8304020927	8304021008	1	1
3	8301161138	8301261131	1	5	45	8304021008	8304021125	1	1
4	8301261207	8301290913	1	3	46	8304021204	8304030929	1	3
5	8301291158	8301300908	1	1	47	8304031226	8304171204	0	27
6	8302021223	8302050919	0	3	48	8304171223	8304200933	0	8
7	8302061122	8302070931	1	2	49	8304201223	8304210920	0	1
8	8302080936	8302081008	0	1	50	8304211215	8304220927	1	8
9	8302081059	8302081213	0	1	51	8304221014	8304221037	1	1
10	8302081213	8302090910	1	2	52	8304221225	8304230904	1	2
11	8302091059	8302121003	1	2	53	8304230932	8304231032	0	1
12	8302121307	8302130939	0	5	54	8304231153	8304240901	1	3
13	8302131220	8302140924	0	4	55	8304240959	8305061037	0	15
14	8302141206	8302150924	1	1	56	8305061109	8305070915	0	1
15	8302161216	8302261004	1	22	57	8305071201	8305101216	0	1
16	8302261004	8302261210	0	2	58	8305181012	8305190914	0	1
17	8302261210	8302261228	0	1	59	8305201152	8305201227	1	1
18	8302261228	8302271015	0	1	60	8305201227	8305211036	0	3
19	8302271312	8302280918	0	1	61	8305211036	8305211044	1	1
20	8302280918	8302280928	1	1	62	8305211044	8305211112	1	2
21	8302280928	8302281032	0	1	63	8305211226	8305241006	0	5
22	8302281136	8302290957	1	1	64	8305241052	8305241130	1	1
23	8302291225	8302300905	1	3	65	8305251140	8305251214	1	1
24	8302301229	8303021002	0	2	66	8305251214	8305260929	0	2
25	8303021222	8303030910	1	2	67	8305260929	8305260950	1	1
26	8303041103	8303050907	1	1	68	8305271103	8305280947	0	1
27	8303061159	8303090904	0	9	69	8305281015	8305281131	0	1
28	8303091229	8303100901	0	3	70	8305281146	8305311040	0	6
29	8303101200	8303110939	1	1	71	8305311224	8306010905	1	1
30	8303161152	8303170911	1	2	72	8306011133	8306011207	1	1
31	8303171129	8303180959	0	2	73	8306011207	8306020910	0	3
32	8303191154	8303200903	1	1	74	8306021229	8306030948	1	1
33	8303200903	8303230945	0	8	75	8306041210	8306070902	0	2
34	8303231217	8303240932	0	3	76	8306070902	8306071146	1	1
35	8303241221	8303250928	1	1	77	8306080901	8306100923	0	1
36	8303251149	8303260908	1	1	78	8306101030	8306101050	1	1
37	8303271208	8303300931	1	6	79	8306101120	8306110936	0	1
38	8303300941	8303301040	1	1	80	8306141207	8306150916	0	1
39	8303301229	8303310909	1	4	81	8306151039	8306151113	1	1
40	8303311114	8304010913	1	4	82	8306161033	8306161055	0	1
41	8304010919	8304010953	0	1	83	8306161203	8306170901	0	1
42	8304010953	8304011021	1	1	84	8306171248	8306181002	1	2

#	Segments Beginning & Ending Time		Trend /Doc		#	Segments Beginning & Ending Time		Trend /Doc	
85	8306181033	8306181102	1	1	129	8309100953	8309111005	0	3
86	8306181216	8306210907	0	5	130	8309111005	8309140944	0	6
87	8306211129	8306220957	0	2	131	8309141131	8309141144	1	1
88	8306221229	8306240946	1	3	132	8309141144	8309151031	0	2
89	8306241113	8306300901	1	9	133	8309151031	8309151226	0	1
90	8306301229	8306310914	0	1	134	8309151226	8309160921	1	1
91	8306311217	8307011045	0	1	135	8309161024	8309161102	1	1
92	8307071026	8307110923	0	2	136	8309161255	8309171003	0	2
93	8307111225	8307120943	0	1	137	8309171129	8309171207	0	1
94	8307140942	8307141045	1	1	138	8309211228	8309221123	1	11
95	8307151149	8307180902	0	1	139	8309221229	8309230928	0	3
96	8307181121	8307190913	1	1	140	8309231108	8309231221	1	1
97	8307201136	8307210953	1	1	141	8309231221	8309240918	0	5
98	8307221127	8307250924	0	4	142	8309241026	8309241136	1	3
99	8307251059	8307260901	1	2	143	8309241136	8309250938	0	3
100	8307261053	8307270936	1	2	144	8309281105	8309291031	0	1
101	8307291058	8308020906	0	2	145	8309291135	8309291314	1	1
102	8308021258	8308030920	1	3	146	8309291314	8309300902	0	5
103	8308031058	8308040923	1	1	147	8309301055	8309301204	0	1
104	8308041216	8308050917	1	4	148	8309301204	8310011035	1	3
105	8308051145	8308060927	1	4	149	8310011120	8310011210	1	1
106	8308061029	8308061206	1	1	150	8310011217	8310020903	0	3
107	8308061206	8308090906	1	6	151	8310021151	8310051006	1	3
108	8308111155	8308120934	0	4	152	8310051219	8310060907	1	3
109	8308120937	8308121013	0	1	153	8310061216	8310070957	1	2
110	8308121224	8308131013	1	1	154	8310071137	8310080932	0	2
111	8308131013	8308131207	1	1	155	8310081308	8310090936	0	3
112	8308131221	8308160921	1	2	156	8310091217	8310121008	0	6
113	8308161209	8308171056	1	3	157	8310121229	8310130943	0	2
114	8308171226	8308180945	1	2	158	8310130943	8310131007	1	1
115	8308191229	8308201021	0	5	159	8310141226	8310151015	1	4
116	8308231228	8308250923	0	1	160	8310151015	8310161012	1	4
117	8308261100	8308270924	1	2	161	8310161135	8310161215	1	1
118	8308271123	8308271208	0	1	162	8310161215	8310190913	0	3
119	8308271229	8308301000	0	4	163	8310191153	8310200910	1	2
120	8308301213	8309010936	1	2	164	8310201214	8310211004	0	1
121	8309011117	8309020919	0	6	165	8310211220	8310220904	1	1
122	8309021210	8309030930	0	2	166	8310220938	8310221051	0	1
123	8309031307	8309040930	0	2	167	8310221132	8311071045	0	51
124	8309041215	8309070929	0	4	168	8311071215	8311110918	0	2
125	8309071219	8309081003	0	4	169	8311111108	8311120918	1	2
126	8309081003	8309081102	1	1	170	8311121136	8311121229	0	1
127	8309081102	8309090901	1	3	171	8311121229	8311130958	1	4
128	8309091137	8309100940	1	6	172	8311130958	8311131214	0	1

#	Segments Beginning & Ending Time		Trend /Doc		#	Segments Beginning & Ending Time		Trend /Doc	
173	8311131214	8311141152	1	6	217	8401301052	8402031039	1	8
174	8311171212	8311180901	1	1	218	8402031254	8402041006	1	4
175	8311201157	8311211013	1	1	219	8402041118	8402050908	0	7
176	8311241210	8311251118	0	4	220	8402051214	8402070943	0	6
177	8311251118	8311261106	0	2	221	8402071222	8402100903	0	1
178	8311261200	8311270912	0	2	222	8402101137	8402101155	1	1
179	8311270919	8311271046	0	1	223	8402101155	8402110924	0	1
180	8311271046	8311281204	0	4	224	8402111229	8402120936	0	2
181	8311281204	8312030936	1	8	225	8402121203	8402130908	0	1
182	8312030952	8312031113	1	1	226	8402141056	8402170918	0	5
183	8312031113	8312040917	0	7	227	8402171002	8402171159	0	2
184	8312041218	8312051130	1	4	228	8402171224	8402180923	0	3
185	8312051155	8312051216	1	1	229	8402181228	8402190924	0	2
186	8312051216	8312080951	1	7	230	8402191139	8402200948	1	4
187	8312081039	8312091103	1	2	231	8402211224	8402240903	0	3
188	8312091103	8312100911	0	2	232	8402241208	8402250919	1	4
189	8312111222	8312120908	0	1	233	8402251153	8402260948	1	4
190	8312121212	8312150947	1	2	234	8402260948	8402261007	0	1
191	8312150947	8312151038	0	1	235	8402261007	8402261024	1	1
192	8312151052	8312151159	0	1	236	8402261059	8402261133	1	2
193	8312151216	8312160922	0	2	237	8402261219	8402270907	0	5
194	8312161059	8312170912	0	2	238	8402281059	8402311019	0	5
195	8312171229	8312180936	0	1	239	8402311124	8403010910	1	6
196	8312181137	8312190956	0	4	240	8403011034	8403011121	0	1
197	8312191245	8312220914	0	2	241	8403011222	8403021124	1	1
198	8312220914	8312221011	0	1	242	8403021200	8403030913	0	7
199	8312221214	8312230944	1	2	243	8403031213	8403040907	0	1
200	8312231212	8312240915	1	1	244	8403041000	8403041047	1	1
201	8312241217	8312250906	0	1	245	8403041202	8403070901	0	7
202	8312251124	8312260908	0	3	246	8403071121	8403071200	1	1
203	8312261308	8401060929	1	3	247	8403071226	8403080953	1	2
204	8401060929	8401071229	1	3	248	8403080953	8403081048	0	1
205	8401071229	8401081048	1	1	249	8403081227	8403091115	0	2
206	8401101227	8401150946	0	5	250	8403091217	8403101004	0	3
207	8401151008	8401151100	0	1	251	8403101201	8403110937	1	3
208	8401171212	8401210925	1	3	252	8403111057	8403111153	0	1
209	8401211001	8401220909	0	3	253	8403111225	8403160918	1	1
210	8401221048	8401230937	1	1	254	8403161229	8403170954	0	5
211	8401241157	8401271013	0	6	255	8403171001	8403171031	1	1
212	8401271049	8401271128	1	1	256	8403171131	8403171217	0	3
213	8401271148	8401271156	1	1	257	8403171217	8403180921	0	5
214	8401271224	8401280924	1	1	258	8403181220	8403210936	1	3
215	8401281156	8401290917	0	2	259	8403210936	8403211013	0	1
216	8401291226	8401300951	1	1	260	8403211206	8403211214	0	1

#	Segments Beginning & Ending Time		Trend /Doc		#	Segments Beginning & Ending Time		Trend /Doc	
261	8403211229	8403220930	0	4	305	8406161201	8406191038	0	7
262	8403221131	8403230955	0	5	306	8406191153	8406200936	1	1
263	8403231138	8403231219	0	3	307	8406200936	8406211042	1	4
264	8403231219	8403240900	0	10	308	8406211059	8406221012	0	3
265	8403241211	8403250909	1	1	309	8406221037	8406230912	1	6
266	8403251248	8403280901	1	1	310	8406230912	8406231003	0	1
267	8403281059	8403291002	0	1	311	8406231117	8406260901	1	20
268	8403291216	8403300900	1	6	312	8406261025	8406270900	0	1
269	8403301220	8403310909	1	4	313	8406271112	8406271148	1	1
270	8403311026	8403311108	0	1	314	8406271225	8406280948	0	1
271	8404011028	8404040908	1	7	315	8406281214	8406300924	0	1
272	8404041128	8404050928	1	4	316	8406301046	8406301150	0	1
273	8404050928	8404060918	1	8	317	8406301150	8406301217	0	1
274	8404061057	8404061129	0	1	318	8406301217	8407020921	1	9
275	8404061131	8404070914	1	4	319	8407020947	8407021031	0	1
276	8404071250	8404080947	1	4	320	8407021247	8407030928	0	5
277	8404081219	8404110922	0	9	321	8407030941	8407031021	0	1
278	8404121001	8404121123	0	1	322	8407031229	8407051102	0	2
279	8404121123	8404130903	1	6	323	8407051104	8407060901	1	2
280	8404131144	8404140931	0	2	324	8407060901	8407100913	1	12
281	8404140931	8404141045	0	1	325	8407111021	8407111116	0	1
282	8404141301	8404151008	0	8	326	8407111206	8407130912	1	5
283	8404151302	8404180923	0	2	327	8407131129	8407131212	1	1
284	8404181200	8404200920	0	1	328	8407171011	8407171151	0	1
285	8404201112	8404210900	0	1	329	8407171151	8407181018	0	1
286	8404211118	8404211134	0	1	330	8407181018	8407190929	1	3
287	8404221137	8404250908	1	10	331	8407190948	8407191027	1	1
288	8404251050	8404251107	1	1	332	8407191043	8407200916	0	3
289	8404251227	8404261019	0	1	333	8407241110	8407241251	0	1
290	8404261028	8404261104	1	1	334	8407241251	8407250919	1	2
291	8404261143	8405290943	0	95	335	8407251130	8407251206	0	1
292	8405291037	8405300908	0	5	336	8407251226	8407260918	1	4
293	8405301034	8405310910	1	1	337	8407261012	8407261046	1	1
294	8405310910	8405310939	1	1	338	8407261046	8407261126	0	2
295	8405311007	8405311036	1	1	339	8407261126	8407270940	1	4
296	8405311120	8406010939	0	7	340	8407271218	8407300912	1	7
297	8406011222	8406021031	0	4	341	8407300944	8407301123	0	4
298	8406021212	8406050903	0	1	342	8407301214	8408011010	0	1
299	8406051115	8406060932	1	2	343	8408020931	8408020947	0	1
300	8406091216	8406120924	0	12	344	8408021126	8408040904	1	3
301	8406121109	8406130917	0	5	345	8408041209	8408070935	0	4
302	8406131203	8406140901	0	2	346	8408071207	8408080915	0	2
303	8406141203	8406150937	1	2	347	8408081008	8408081135	0	1
304	8406150958	8406160921	1	2	348	8408081224	8408090915	0	2

#	Segments Beginning & Ending Time		Trend /Doc		#	Segments Beginning & Ending Time		Trend /Doc	
349	8408091225	8408161049	1	12	393	8410191125	8410200907	1	5
350	8408161049	8409011257	1	52	394	8410201109	8410240958	1	3
351	8409011257	8409020949	1	1	395	8410241107	8410250913	1	4
352	8409021207	8409050901	1	8	396	8410251132	8410271152	1	6
353	8409050931	8409051200	0	1	397	8410271212	8410281058	0	1
354	8409051200	8409060901	1	5	398	8410281058	8411031215	1	8
355	8409061205	8409081006	1	1	399	8411031215	8411040916	1	8
356	8409090932	8409091047	1	1	400	8411040916	8411091223	0	6
357	8409091207	8409120919	1	3	401	8411091223	8411111109	1	9
358	8409121141	8409130912	1	1	402	8411111109	8411121003	0	1
359	8409130912	8409130942	1	1	403	8411121003	8411150955	1	3
360	8409131019	8409140946	0	3	404	8411150955	8411160923	0	5
361	8409141037	8409141101	1	1	405	8411161018	8411161113	0	1
362	8409141208	8409150905	1	2	406	8411161149	8411170905	1	1
363	8409151103	8409151128	0	1	407	8411170905	8411180902	1	3
364	8409151128	8409190949	1	11	408	8411180902	8411181006	0	1
365	8409190949	8409191014	0	1	409	8411181006	8411181149	1	2
366	8409191151	8409200911	1	3	410	8411181216	8411230950	1	3
367	8409200929	8409201042	1	1	411	8411230950	8411231131	0	2
368	8409201042	8409201121	0	1	412	8411241126	8411250901	1	2
369	8409201121	8409201227	0	1	413	8411250901	8411300901	0	16
370	8409201227	8409210940	1	6	414	8411300901	8412081018	0	25
371	8409211143	8409220948	1	3	415	8412091058	8412091227	0	1
372	8409221120	8409221224	0	1	416	8412091227	8412101013	0	2
373	8409221224	8409230946	1	6	417	8412101013	8412131134	0	3
374	8409231212	8409260922	1	6	418	8412131134	8412141143	0	1
375	8409261109	8409270918	0	5	419	8412141148	8412150914	0	4
376	8409271017	8409271050	1	1	420	8412161229	8412171206	0	9
377	8409271144	8409271229	0	1	421	8412171206	8412200942	0	6
378	8409271229	8409280948	1	6	422	8412200942	8412201042	1	1
379	8409281054	8409281132	1	1	423	8412201042	8412210928	1	2
380	8409281226	8409301031	1	15	424	8412211227	8412220917	0	2
381	8409301132	8410030957	1	8	425	8412220917	8412221034	0	1
382	8410031035	8410041112	0	4	426	8412221217	8412230904	1	1
383	8410041112	8410061141	0	16	427	8412231147	8412240917	1	1
384	8410061149	8410071030	0	4	428	8412241225	8412241238	0	1
385	8410071045	8410071118	0	1	429	8412241238	8412280907	1	5
386	8410071156	8410100927	1	4	**Total # of Aligned Documents: 1516**				
387	8410100927	8410111229	1	4					
388	8410111229	8410130906	1	6	**Total # of Aligned Segments: 429**				
389	8410130906	8410140910	1	6					
390	8410140910	8410171041	1	5	**# of News Aligned to Rise Trends: 717**				
391	8410171142	8410180901	1	2					
392	8410181031	8410190907	1	2	**# of News Aligned to Drop Trends: 799**				