Stock Market Random Forest-Text Mining (SMRF-TM) Approach to Analyse Critical Indicators of Stock Market Movements

MAZEN NABIL ELAGAMY

A thesis submitted in partial fulfilment of the requirement of Staffordshire University for the degree of Doctor of Philosophy

November 2017

Abstract

The Stock Market is a significant sector of a country's economy and has a crucial role in the growth of commerce and industry. Hence, discovering efficient ways to analyse and visualise stock market data is considered a significant issue in modern finance. The use of data mining techniques to predict stock market movements has been extensively studied using historical market prices but such approaches are constrained to make assessments within the scope of existing information, and thus they are not able to model any random behaviour of the stock market or identify the causes behind events. One area of limited success in stock market prediction comes from textual data, which is a rich source of information. Analysing textual data related to the Stock Market may provide better understanding of random behaviours of the market.

Text Mining combined with the Random Forest algorithm offers a novel approach to the study of critical indicators, which contribute to the prediction of stock market abnormal movements. In this thesis, a Stock Market Random Forest-Text Mining system (SMRF-TM) is developed and is used to mine the critical indicators related to the 2009 Dubai stock market debt standstill. Random forest and expectation maximisation are applied to classify the extracted features into a set of meaningful and semantic classes, thus extending current approaches from three to eight classes: critical down, down, neutral, up, critical up, economic, social and political. The study demonstrates that Random Forest has outperformed other classifiers and has achieved the best accuracy in classifying the bigram features extracted from the corpus.

Acknowledgments

To my guardian angels, my late parents Nabil Elagamy and Nayera Elgamal: no words can express how grateful I am for everything they did for me, I owe it all to them.

I would like to express my sincere gratitude to my supervisors Prof. Bernadette Sharp, Dr. Clare Stanier, Prof. Yasser Elsonbaty and Prof. Mohamed Abo Elnasr for the continuous support in my PhD study and related research, for their patience, motivation, and tremendous knowledge. Their guidance throughout the research and writing of this thesis was a great support for me. I would not have imagined working with a better advisors and mentors for my PhD study.

In addition, I would like to express my deep gratefulness to the domain experts for their time, guidance through the research, insightful comments and valuable feedbacks, which incensed me to modify and expand my research from various perspectives.

Last but by no means least, I would like to thank my family: my wife, my children, my brother and my sister in law for supporting me spiritually throughout the research and writing this thesis.

List of Publications

Elagamy, M. N., Stanier, C. and Sharp, B. (2018). Text Mining Approach to Analyse Stock Market Movement. In *the 3rd International Conference on Advanced Machine Learning Technologies and Applications* (AMLTA). Accepted and to be presented in Cairo in February 2018.

Elagamy, M. N., Stanier, C. and Sharp, B. (2018). SMRF-TM Approach to Analyse Critical Indicators of Stock Market Movements and classify the related news articles. Submitted to the *2nd International Conference on Natural Language and Speech Processing* (ICNLSP), 25-26 April 2018, Algeria.

Elagamy, M. N., Stanier, C. and Sharp, B. (2018). Mining Critical Indicators of 2009 Dubai Stock Market Debt Standstill. In Hassanien A. E. (ed.) *Machine Learning Paradigms: Theory and Applications* Springer, Cairo: Springer Studies in Computational Intelligence (forthcoming publication).

Table of Contents

List of Figures

List of Tables

List of Abbreviations

SM	Stock Market
DM	Data Mining
TM	Text Mining
NLP	Natural Language Processing
KD	Knowledge Discovery
ML	Machine Learning
DT	Decision Tree
RF	Random Forest
SVM	Support Vector Machine
EM	Expectation Maximisation
IR	Information Retrieval
IE	Information Extraction
FE	Feature Extraction
SMRF-TM	Stock Market Random Forest-Text Mining

Chapter 1 Introduction

1.1 Knowledge Discovery and the Study of Stock Markets

Knowledge Discovery (KD) has become one of the most important fields in the information industry due to the increasing amount of data available for analysis and trend discovery. The knowledge extracted from this data is used for different business and financial applications such as production control, stock markets analysis, portfolio management and design of interpretable trading rules and discovering money laundering schemes using decision rules and relational data mining methodology. Data Mining (DM) is a subfield of knowledge discovery and can be defined as the process of extracting hidden patterns and knowledge from large amounts of *structured* data. Specialised data mining tools are able to find patterns in large amounts of structured databases and to analyse significant relationships, which exist only when several dimensions are viewed at the same time (Han & Kamber 2006). Text Mining (TM), which is the focus of this research, is another subfield of knowledge discovery. It is an exciting area of computer science research, which tries to address the crisis of information overload by combining techniques from data mining, machine learning, natural language processing, information retrieval and extraction, and knowledge management. It is a multidisciplinary field as it involves the retrieval and pre-processing of document collections, language analysis and the intermediate representations of significant concepts extracted from the documents, data mining techniques to analyse these intermediate representations, and the visualisation of the generated results (Feldman & Sanger 2007, Tan 1999). Text mining can be defined as the process of extracting important and non-trivial knowledge from *unstructured* textual data. Consequently, text mining is considered to be more complex than data mining as it deals with unstructured, fuzzy and ambiguous textual data. It is also believed to

have a more powerful commercial value than that of data mining since the textual form is the utmost common form of storing information.

The application domain of this research is the Stock Market (SM) also known as equity market or share market; it is the market where shares of public listed companies are issued and traded. The stock market makes it possible to grow small initial sums of money into large ones without taking the risk of starting a new business. It is a very important sector of the economy of a country as it plays a crucial role in the growth of commerce and the industry of the country and it is also believed to be one of the most significant sectors of a free market economy, as it provides companies with access to capital in exchange for giving investors a slice of ownership in the company. When a stock market is rising this is a good indication for a developing industrial sector and a growing economy of the country, so the central banks of all the countries and the governments carefully monitor the stock market on a continuous basis. In addition, stock market is the main source for any company to raise funds for business expansions (Cheema *et al.* 2008).

The increasing importance of the stock markets and their direct influence on the economy were the main reasons for deciding to study and analyse stock market crashes as the application domain of this research. The 2009 Dubai stock market debt standstill was chosen as the specific application domain for this research. There were two reasons for choosing the Dubai debt standstill, firstly data collection and secondly validation. Consideration was initially given to using other stock market crises such as the 1929 Wall Street crash, the 1973-1974 United Kingdom stock market crash, the 1998 Russian financial crisis, and the Chinese stock bubble of 2007. However, it proved very difficult to collect enough textual data (financial news) relevant to these crashes, which was suitable for analysis. This was especially the case for very old crashes. Secondly, the nature of the research required the use of the financial experts who could qualitatively validate the

research results. The financial experts who were available to validate this research have expertise in the Middle East stock markets and this meant that the Dubai crisis was a suitable domain against which to validate the Stock Market Random Forest-Text Mining system (SMRF-TM) developed in this thesis.

The use of data mining techniques to analyse stock markets has been extensively studied using structured data like past prices, historical earnings, or dividends. However, text mining approaches are comparatively rare due to the difficulty of extracting relevant information from unstructured data. As Patel *et al.* (2015) claim, stocks behave randomly. Furthermore, Schumaker *et al.* (2012) and Nikfarjam *et al.* (2010) explain that the application of data mining to the analysis of stock market data using current approaches may not be sufficient to model and justify any random behaviour of the market based only on quantitative data such as the values of stocks and historical market prices. This suggests that if researchers focus on the impact of un-quantifiable events on the market, which can be extracted from related news articles, they may be able to justify the random behaviour of the market and to enhance the analysis performance. Drury (2013) stated that there are huge amounts of free news and financial data, which are believed to contain rich information known as “alpha”. Alpha is considered to be valuable, non-trivial and rich information embedded in textual data, which can be very useful for the purpose of analysis. The hypothesis of his research is that text mining approaches can be applied to enhance the performance of current trading systems’ strategies if the “alpha” embedded in financial news is used to support the prediction of stock market share price movement directions.

Consequently, discovering efficient ways to analyse and visualise stock market features is considered a significant issue in modern finance not only to be able to give individuals, institutions or countries useful information about the market behaviour for investment decisions, but also because stock markets can

dramatically affect important financial and economic factors (Farmer 2015, Hajizadeh *et al.* 2010, Mishkin & White 2002). In order to study such effects Mishkin and White (2002) examined fifteen episodes of stock market crashes in the United States in the twentieth century highlighting the impact of the crashes of 1929 and 1987 and the resulted stress on the financial system. They demonstrated how the crashes of 1907, 1930-33, 1937 and 1973-74 were associated with large increases in spreads causing severe financial distress. Farmer (2015) investigated the relationship between stock market and unemployment rate. The results of his research showed that the stock markets' movements is responsible for the unemployment rate and that over a seventy year period the relationship between stock markets' movements and unemployment rate changes had a stable structure. He also showed that the drop in the stock market, which occurred in autumn of 2008, was one of the main reasons for the magnitude of the recession, which followed.

1.2 Motivation for Analysing Stock Market Movement

Even though the ability to analyse stock market movement has been a source of interest for many researchers, a satisfactory method for analysing stock price movement with acceptable performance has not yet been developed. The cause of the difficulty in the analyses of the stock market is the complexities associated with market dynamics where parameters are not fully defined and are constantly shuffling (Schumaker *et al.* 2012, Schumaker & Chen 2009).

In recent years, there has been an increase of interest in quantitative funds, which automatically shift through numeric financial data and issue stock recommendations (Schumaker & Chen 2009). While these systems are based on proprietary technology, they differ in the amount of trading control they have, ranging from simple stock recommenders to trade executors. Using historical market data and complex mathematical models, these methods are constrained to

make assessments within the scope of existing information. This weakness means that they are unable to react to unexpected events falling outside of historical norms (Schumaker *et al.* 2012, Nikfarjam *et al.* 2010).

The use of data mining techniques, such as classification and regression trees, chi-squared automatic induction, neural networks and genetic algorithms, to predict the stock market has been extensively studied using structured data (Mittermayer & Knolmayer 2006). The stock market is a chaotic, dynamic and complicated system, which is considered to be one of the core financial tasks for data mining (Nakhaeizadeh *et al.* 2002). The main reasons for researchers to use data mining techniques in the prediction of financial markets are their need to forecast a multidimensional time series, which contain a very high level of noise, accomplish an integrated multidimensional forecast to sustain certain efficiency criteria with a reasonable prediction accuracy, consolidate flow of textual data for forecasting models as input data and also to be capable to justify the forecast and the forecasting model as well (Hajizadeh *et al.* 2010). But there is still a major problem for better predictions in approaches just based on historical market prices, which is the ability to model any random behaviour of the market. Random behaviour is very difficult to justify because quantitative data solely cannot explain any random behaviour of the stock market (Nikfarjam *et al.* 2010). Also, data mining analysis makes use only of quantifiable information in terms of charts or numeric time series, which only describe the event but not their causes (Wuthrich *et al*. 1998).

The issues with textual data are considered to be one of the main reasons for the limited success in stock market analysis. Textual data such as news reports and economical articles are qualitative data, which must be translated to numeric form before many computational systems can process it. However, they are an important source of information about stock market and their analysis may provide a better understanding of random behaviour of the market, which is difficult to explain by

focusing solely on statistical data (Schumaker *et al.* 2012). For this reason, we collected the relevant documents related to Dubai debt standstill for the proposed text mining study since the focus of this research is on one specific domain of study, namely the Dubai debt standstill dated 27 November 2009. However, relying solely on the analysis of these textual data has some limitations. The importance of news events can only be evaluated at a later time, and experts may have different opinions and interpretations of the events. Also, the lack of sufficient and clear information about relationships between decision variables and outcomes always make experts and investors lapse into making relatively less rational decisions in financial market. This problem becomes worse when decision makers are confronted with large amounts of information (Wang *et al.* 2011). However, textual data does provide a wealth of data but many fund managers have been unable to fully capitalise on this because information, which is implicit in the data for the purpose of investment is not easy to discern (Kannan *et al.* 2010). The key issue is the necessity to use the user's specification to label historical documents for training and classifying. Textual information is complex and rich. Whilst tables with financial data indicate how well a company has achieved, the linguistic structure and written style of the text may reveal more about its strategy and future performance (Kloptchenko *et al.* 2002). The use of textual data relies heavily on human analysis in order to achieve a better analysis of stock market price movements. Unlike numerical and fixed field data, it cannot be analysed by standard statistical data mining method (Nasukawa & Nagano 2001). Even though text mining is expected to play an important role in designing strategies for the analysis of market behaviour, to the best of our knowledge this is still a relatively new field and there is a lack of research on the use of text mining to understand the causes of stock market movements and improve the analysis of stock market (Nikfarjam *et al.* 2010).

1.3 Research Questions

This research is designed to address the following questions:

(a) Can the random forest algorithm support the identification of the critical indicators, which affect the stock market movements?

(b) Can the extension from three to five classes of indicators enhance the classification performance?

(c) Can the expectation maximisation algorithm be used to examine the reasons behind Dubai's stock market movements by classifying these indicators into their semantic attributes (i.e. economic, social or political)?

1.4 Research Aim and Objectives

The main focus of this research is the application of text mining to investigate and analyse textual information (news and historical documents), and of random forest to identify the critical indicators, which contribute to the understanding of stock movements. To achieve this aim the following objectives were developed:

- To review current text mining methods and approaches to the analysis of stock market domain.
- To focus on one specific domain of study, namely the Dubai debt standstill dated 27 November 2009, and to collect the relevant documents related to Dubai debt standstill for the proposed text mining study.
- To review current feature extraction methods and implement the best approach to extract key terms, which can best capture critical indicators related to stock market.
- To implement the random forest algorithm to analyse the extracted critical indicators.

- To cluster these indicators using expectation maximisation algorithm according to their semantic categories, and validate the extracted stock market critical indicators against the experts' critical indicators.
- To refine the novel approach based on the experts' validation.
- To use cross-validation in order to evaluate the method and the final outcomes of the random forest.

1.5 Research Methodologies

1.5.1 Research Philosophy

Any research should be based on some underlying epistemology. Epistemology refers to the assumptions about what constitutes valid research and which methods are appropriate for the research domain. Hence, it is important to know these assumptions to be able to conduct and evaluate a research (Hirschheim 1992).

In this research, the suggestion of Orlikowski and Baroudi (1991) and Chua (1986) is adopted, which is defining three classes for underlying research epistemology: positivist, interpretivist and critical. Depending on the underlying philosophical assumptions of the researcher the research can belong to any class of these three classes. Positivist researches usually try to test a theory in order to enhance the understanding of phenomena through the assumption that reality is objectively given and can be defined by measurable variables, which is independent of the researcher (Orlikowski & Baroudi 1991). On the other hand, interpretive researchers generally try to understand phenomena through the meanings, which people assign to them. Hence, interpretive researchers are concerned with the decisions made by humans as the situation occurs assuming that reality can only be accessed through social constructions such as language, shared meanings and awareness (Kaplan & Maxwell 1994, Orlikowski & Baroudi 1991).

This research is an interpretive in-depth case study research, which aims to analyse Dubai's stock market debt standstill occurred in 2009 through applying text mining methods to study the critical indicators, which contribute to the prediction of abnormal stock movements. It can also be considered as a positivist comparative research as it deploys quantitative approaches to compare the results yielded by applying Random Forest classifier against another set of classifiers such as ADTree, J48, J48graft, Decision Stump, Random Tree, Bayes Net, Bagging, Rotation Forest and Decision Table.

1.5.2 Research Approach

There are two main approaches for research, quantitative and qualitative approaches, which are associated with the positivist paradigm (quantitative) and the interpretive paradigm (qualitative). Quantitative based research consists of studies in which the data can be analysed in terms of numbers such as survey methods, laboratory experiments and mathematical modelling and was developed from the natural sciences. Qualitative research involves the use of qualitative data, such as interviews, documents, and participant observation data in order to understand and explain social phenomena. Quantitative approach developed in the social sciences to enable researchers to study social and cultural phenomena. In addition, quantitative approaches use deductive logic facilitating the ability to choose concepts, variables and hypotheses before the study begins. On the other hand, qualitative approaches use inductive logic so categories emerge from the informants and lead to patterns or theories, which help to explain a phenomenon (Myers 1997). Quantitative and qualitative approaches are not mutually exclusive and researchers may use both approaches in what is termed a 'mixed methods' approach. This research combines quantitative data based on analysis of stock market movements with qualitative data reflecting views and opinions. We therefore

adopt a mixed method approach, which is sometimes referred to as triangulation approach (Mingers 2001, Gable 1994, Markus 1994).

1.5.3 Research Case Study and Design

The case study of this research is Dubai's stock market debt standstill 2009. Dubai is one emirate out of the seven United Arab Emirates, which have different ruling families and budgets. Dubai's economy depends on trade, ports, services and finance. When the international financial crisis of 2007-2010 occurred the real estate market in Dubai dramatically declined in November 2009 after a six-year boom. Dubai had about $80bn of debts of which $60bn belonged to *Dubai World*, the state-owned holding company, which was responsible for triggering the crisis in Dubai. Consequently, the Dubai government asked all the financing providers for *Dubai World* to standstill and extend maturities for six months. (www.telegraph.co.uk).

Textual data (financial news) relevant to the case study for the purpose of analysis was collected through a formal subscription to the official web site of the Financial Times. A total of 544 financial news articles concerning Dubai's stock market, published in the period between 2008 till 2012, were retrieved. The retrieved data is used to quantitatively validate and analyse the proposed approach using k-fold cross validation and text mining techniques such as, term frequency-inverse document frequency, random forest, and expectation maximisation in order to identify the critical indicators, which can seriously affect the prediction performance of stock market movements. Then a qualitative validation of the results yielded was carried out using financial experts.

1.6 Contributions to Knowledge

The major contributions of this research include the following:

(i) The application of random forest to a new domain, which is the analysis of stock market textual data using text mining techniques.

(ii) The extension of the classes used to classify the extracted features and the news articles from three classes (good, bad or neutral) to five meaningful classes (critical down, down, neutral, up and critical up).

(iii) The application of the expectation maximisation clustering technique to cluster the classified features according to their semantic attributes (economic, social or political).

(iv) The developed SMRF-TM system is able not only to classify the features/articles according to the predicted influence they have on Dubai's stock market movements, but also able to describe the causes behind these classifications.

1.7 Ethical Basis

This research project was conducted in full compliance with the ethical regulations of Staffordshire University and the British Computing Society code of conduct. The articles, which provided the texts explored through the random forest algorithm were all in the public domain and accessing these texts had no ethical implications. This research project respected the confidentiality and anonymity of the experts, and ensured that their participation is voluntarily. They were fully informed of the aim of this research project and they have rights to withdraw from the study at any stage. This thesis did not seek any participation from children, people with communication or learning difficulties, patients, people in custody, people who can be considered vulnerable or people engaged in illegal activities. Finally, this research project has adopted appropriate ethical and professional standards and responsibilities in its publications; all external sources of information are acknowledged and attributed professionally. A sample of the consent form is found in appendix D.

1.8 Thesis Structure

This thesis consists of six chapters. Chapter one has introduced the background themes to this research: knowledge discovery and stock markets, the motivation of the research and the research methodologies adopted to achieve the research aims and objectives. Chapter two reviews the literature related to stock market, data mining and text mining. As the approach of this research is based on text mining, chapter three discusses the principles and stages of text mining. Chapter four describes the implementation stages of the proposed Stock Market Random Forest-Text Mining (SMRF-TM) approach and discusses the experimental works and results. Chapter five explains the validation and evaluation techniques used in the proposed (SMRF-TM) approach. Chapter six summarises the research approach, discusses the challenges and limitations encountered through the research and finally proposes some recommendations for the future work.

Chapter 2 Literature Review

2.1 Introduction

Financial data analysis has traditionally dealt with large volumes of structured data reflecting economic performance. However the behaviour of the market is dictated by contemporary local and global events, such as domestic and international news, financial and government reports and natural disasters etc., which are not captured in the statistical data (Wu *et al.* 2014, Gómez *et al.* 2001). Consequently, we need first to show that trading on information “alpha” embedded in financial news can attain a profitable trading approach as markets react to news stories. This can be shown through a shallow economical literature review followed by a deeper literature review on existing Stock Market (SM) prediction systems deploying Data Mining (DM) and Text Mining (TM) techniques.

2.2 Related Work

2.2.1 Economical Aspects for Predicting Stock Market Movements

The prediction of stock markets movements is significant for economical researchers from more than one perspective. Empirically, studying stock markets movements reveal information about stock markets’ driving factors. From a theoretical point of view, this can be viewed as assessments of existing asset pricing theories. Hence, there are extensive studies in financial economics, which addressed this issue (Pönkä 2017).

Niederhoffer (1971) was the first one who used news information “alpha” in order to enhance the performance of a real-world trading approach. This was done by classifying stories in the print media into 19 different categories to express a polarity

scale from encouraging to discouraging. He was capable to produce a reasonable trading approach.

Schuster (2003) shows that it is not a must that all events aggravate a reaction, through reviewing huge events and the reaction of the S&P market index to those events (Robbani & Anantharaman 2004, Culter *et al.* 1991). Only unexpected events cause ultimate effect while expected events tend to aggravate no reaction at all. But as financial news published in the mass media should be unexpected in order to attract readers and be interesting for publishing so the publication of events in the mass media is expected to lead to stock market reaction (Drury 2013, Schuster 2003, Bomfim 2000, McManus 1988).

Davis *et al.* (2006) state that sentiment in news can indicate future performance because there is a correlation between language usage and future performance, which can be shown by the market response to optimistic and pessimistic language usage in earnings press releases. By analysing the writing style of company reports, Henry (2006) found that diversity in writing style from pessimistic to optimistic could indicate company's future expectations. Showing that more definite predictions of market response can be achieved by using predictor variables, which capture verbal content and writing style of earnings press releases. Later, Tetlock *et al.* (2008) found that fraction of negative words in firm specific news stories forecasts low firm earnings and that firms' stock prices under react to the information embedded in negative words (Drury 2013). In addition, Ravenpack Company has produced a news analytic system, which demonstrated that there were correlations between sentiment in news and the following two weeks returns in the Eurstoxx and Dow Jones market indexes (Drury 2013, Hafez 2009). Consequently, sentiment analysis systems, which explore emotions and feelings expressed in natural language texts, can be used to support the extraction of important information embedded in textual data (Glucksberg 2008).

2.2.1.1 Markets under/over Reaction to an Event

De Bondt and Thaler (1985) claimed that markets over-react or under-react to an event and that a subsequent price movements in the reverse direction will correct movements in stock prices where the larger the initial price movement the higher will be the subsequent price movement. Consequently, Hong and Stein (1999) proposed the unified theory of under-reaction, momentum trading, and over-reaction in asset markets based on the idea of gradual diffusion of information among investors, which causes prices to under-react in the short run, making it possible for momentum traders to profit from trend chasing. This can be illustrated with reference to the inaccurate information about the United Airlines bankruptcy published in September 2008. The bankruptcy news lowered the share price and when the story was corrected the share price returned back to normal. Using the De Bondt's and Thaler's hypothesis, a trading approach would have bought at the lowest price knowing that this price fall would be followed by a subsequent market correction. The inaccurate information also had a negative effect on some other major airlines (American Airlines, Continental Airlines, Delta Airlines and U.S. Airways) (Carvalhob *et al.* 2011).

2.2.1.2 Spill over Effect

When linked economical actors are affected by forecasts of one economical actor this effect is known as spill over. An example presented by Drury (2013) to emphasis the idea that news stories may affect other economical actors, which are not specifically quoted in the news story was reduction of the credit status of Portugal, which directly affected the cost of government debt for Spain, Italy and Ireland. Hafez (2010) states that since news has influence on exposure and covariance of stocks there are spill over effects from news releases. Mitra and Mitra (2011) and Hafez (2010) assert that the industry index price can be affected by company specific news events. This is because an event regarding a single

company can affect many other companies within the same sector leading to this being reflected in the industry index price.

2.2.1.3 The Effect of News on Stock Prices from the Economical Aspect

Drury (2013) summarised the effect of news on stock prices from the economical aspect into the following four hypotheses: (1) an instant reaction in the stock market can be initiated by events only when the events have economic consequences, (2) expected or insignificant news stories are filtered by the mass media, (3) market reaction to events may be for a shorter period than the market reaction to sentiment information and (4) it is not a must that a company is specifically mentioned in the news text to affect the company's share prices.

2.2.2 Data Mining

The application of data mining techniques for financial markets prediction and classification is considered a very productive research area (Kirkos *et al.* 2007). The nature of financial data, which is a multidimensional time series containing a very high level of noise, is the main reason for researchers to employ data mining techniques in the prediction of financial markets. The use of data mining techniques allows the researchers to accomplish an integrated multidimensional forecast in order to sustain a certain efficiency criteria with a reasonable prediction accuracy and to be capable of justifying the forecast and the forecasting model as well (Hajizadeh *et al.* 2010).

The stock market is a typical example for such financial markets, which continuously produces a huge amount of data such as bids, buys and puts (Wu *et al.* 2014). The stock market is a chaotic, dynamic and complicated system, which is considered to be one of the core financial tasks for data mining (Nakhaeizadeh *et al.* 2002). So, to examine comparable behaviour of traded stock prices Basalto *et al.* (2005) applied a pair wise clustering approach to the analysis of the Dow Jones

index companies in order to understand the underlying dynamics, which rules the companies' stock prices. They employed the chaotic map-clustering algorithm, where a map was identified for each company and the correlation coefficients of the financial time series were associated to the coupling strengths between maps. The simulation of the chaotic map dynamics showed that the companies within the same industrial branch are often grouped together. Then the identification clusters of companies of a given stock market index can be expressed in the portfolio optimisation strategies. A stock trading method, which combines the filter rule and the Decision Tree (DT) techniques was presented by Wu *et al.* (2006), to help decide which stocks to buy and the right timing for buying it as this is a very important issue for investors in stock market domain. They used the filter rule to generate candidate-trading points then these points were clustered and screened by the decision tree algorithm. Using Taiwan and NASDAQ stock markets their experimental results showed that their method is distinct in consolidating future information into criteria for clustering the trading points and that it outperformed both the filter rule and all the previous literature, which applied such a combination technique. Also, a data mining method was designed to incorporate attribute-oriented induction, Information Gain (IG) and decision tree, which is suitable for pre-processing financial data and establishing a decision tree model to predict financial distress for listed companies. Design is based on the financial ratios attributes and one class attribute by adopting an entropy-based discretisation method. The experimental results with 35 financial ratios and 135 pairs of listed companies as initial samples showed satisfying results, testifying the feasibility and validity of the proposed data mining method (Sun & Li 2008). Wang *et al.* (2011) proposed an ontology based data mining framework, which specifically provided an ontology method for processing news data into classes of events to discover actuarial relationships between various kinds of news and market movements in term of price trends, volume changes and similar elements. The reasoning output of the expert

system was used to build a Bayesian network highlighting the dependence relationships between the stocks and possibly significant news, demonstrating the significance order of each kind of news on certain financial instrument trading activity to experts and investors.

A company's financial distress does not only affect the interests of the enterprise and the staff but it also has a negative effect on the investors and the entire related economical sector in the country. Consequently, Geng *et al.* (2015) designed an early financial crisis warning system for listed companies in China. This was done through studying the financial distress phenomenon for 107 Chinese companies, which were labelled by the Shanghai Stock Exchange and the Shenzhen Stock Exchange as "special treatment" from 2001 till 2008. They deployed data mining techniques to build their models according to 31 financial indicators and three time windows. The results of their research showed that neural networks outperformed the other classifiers, which are support vector machine, decision tree and an ensemble of multiple classifiers using majority votes. In addition, Salehi *et al.* (2016) compared the performance of four different data mining techniques, which are support vector machine, artificial neural network, k-nearest neighbour and naïve Bayesian classifiers in order to predict corporate financial distress using accounting data of the Iranian firms for two years prior to financial distress. The results of their research showed that the artificial neural network outperformed the other data mining techniques.

There are extensive studies on applying data mining techniques to predict stock market movements using structured data such as historical market prices. However, these data mining approaches are constrained to make assessments within the scope of existing information, because they only analyse the quantifiable information embedded in charts or numeric time series. Such quantifiable information can only be used to describe the event but not the causes behind such events (Mittermayer & Knolmayer 2006, Wuthrich *et al*. 1998). Consequently, the

lack of capability to describe the causes for events led to the inability to model any random behaviour of the market, which is considered as a major barrier to better predictions in approaches just based on historical market prices (Nikfarjam *et al.* 2010, Mittermayer & Knolmayer 2006). One area of limited success in stock market prediction comes from textual data, which is a rich source of information and analysing it may provide better understanding of random behaviours of the market.

2.2.3 Text Mining

Text mining is the discovery of new, previously unknown information, by automatically extracting information from different resources for textual data. The process of text mining encompasses the following major steps: Information Retrieval (IR), Information Extraction (IE) and data mining (Ghosh *et al.* 2012, Ananiadou *et al.* 2006). Many researchers have explored the field of text mining to understand the causes of stock market movements and improve the prediction accuracy of stock market movements. Whilst tables with financial data indicate how well a company has achieved, the linguistic structure and written style of the text may tell more about its strategy and future performance (Kloptchenko *et al.* 2002). To examine the importance of text analysis for stock price movement prediction, Lee *et al.* (2014) produced a text mining prediction system to forecasts companies' stock price changes (down, stay or up) influenced by financial events reported in 8-k documents. Their results showed that textual analysis enhanced the prediction accuracy around 10% over a powerful baseline, which only deploys data mining techniques to analyse numeric data. This indicates that textual data such as news, financial reports and economical articles are an important source of information about stock market and their analysis may provide a better understanding of any random behaviour of the market, which is difficult to be justified by focusing solely on historical and statistical data. Consequently, text mining is expected to play an important role in designing strategies for prediction of stock market behaviour.

Hence, text mining is the focus of this research aimed at demonstrating its potential and valuable contribution to stock market crashes analysis, which is an important event of today's global economy.

Most textual sources used by text mining researchers for market prediction include financial journals and news such the Wall Street Journal, Financial Times, Reuters, Dow Jones, Bloomberg and even Yahoo Finance, and often the analysis is focused on the news text or the news headlines (Nassirtoussi *et al.* 2014). The literature review reveals two main text mining approaches are adopted in the analysis, prediction or mining of stock market features: (i) machine learning such as Support Vector Machine, decision rules/trees, regression algorithms, naïve Bayes, and (ii) natural language processing algorithms (Gonçalves *et al.* 2013). Machine learning algorithms give computers the ability to learn without being explicitly programmed by involving a set of data to train the algorithm and using another set of data to test the generated predictions. The natural language processing approach involves lexical, syntactic, semantic and pragmatic analysis of unstructured texts (Gonçalves *et al.* 2013).

Gómez *et al*. (2001) used the analysis of the news as a means to elicit the interaction and influence of social interests on their behaviour. They used simple statistical representations of the news reports and statistical measures for analysis and discovery of useful trends. Mahajan *et al.* (2008) employed text mining to analyse financial news articles and reports in conjunction with time-series market data in order to explain the causes for poor performance or a sudden upturn in the market. They proposed a text mining system, which analyses financial news related to the Indian stock market in order to identify the major events, which have impact on the stock market and to design strategies for predicting the market. The events have been studied using Latent Dirichlet Allocation (LDA) based on topic extraction mechanism. The study carried out by (Nikfarjam *et al.* 2010) reveals that automatic

text classification techniques are commonly used in analysing incoming news, and in some cases researchers make use of historical market prices data related to stock price to improve the accuracy of their prediction, thus combining data and text mining algorithms. Such predictive systems consist of three main components: classifier input generation, classification and finally news labelling. Ming *et al.* (2014) propose a unified latent space model to characterise the "co-movements" between stock prices and news articles and to predict the closing stock prices on the same day; their algorithm is based on the analysis of daily articles from Wall Street journal. Sun *et al.* (2016) predict the stock market based on textual information from user-generated micro-blogs using the latent space model to correlate the movements of both stock prices and social media content. Kim *et al.* (2014) apply natural language processing to analyse economic news articles of a media company to categorise and extract the sentiments and opinions expressed by the writers. Their aim is to identify the correlation between news and stock market fluctuations. Ali and Theodoulidis (2014) adopted a linguistic based text mining approach demonstrating how text mining could be integrated with the financial fraud ontology to improve the efficiency and effectiveness of extracting financial concepts. Schumaker and Chen (2009) examined a predictive machine learning approach to analyse financial news articles and stock quotes covering the S&P 500 stock market index during a five weeks period using a set of linguistic textual representations, including bag of words, noun phrases, and named entities approaches to estimate a discrete stock price twenty minutes after a news article was released. Using Support Vector Machine (SVM) derivative tailored to discrete numeric prediction and models they showed that their model had the best performance in closeness to the actual future stock price and the highest return using a simulated trading engine. They have also concluded that a proper noun scheme performs better than bag of words in their metrics. Kannan *et al.* (2010) discussed various techniques (e.g. typical price, relative strength index and moving

average) to predict whether future closing stock price will increase or decrease and to investigate various global events and their influence on predicting stock markets. Nikfarjam *et al.* (2010) considered three market aspects, such as input data, predictive goal and prediction horizon, to predict the price and volatility of the market based on the new content. Using machine learning techniques, they labelled the news and classified them to investigate the impact of financial news on stock market prediction. Similarly, Kaya and Karsligil (2010) classified financial news articles into positive or negative according to their effects on stock price based on price changes to label the articles and using support vector machine. Nassirtoussi *et al.* (2014) summarised studies, which are concerned with weighting text for predicting stocks price movements. In addition, they also reviewed the performance of various text mining methods applied using different text sources and they showed that most textual sources used by text mining researchers for market prediction include financial journals and news such the Wall Street Journal, Financial Times, Reuters, Dow Jones, Bloomberg and even Yahoo Finance, and often the analysis is focused on the news text or the news headlines. Bollen *et al.* (2011) inspected whether public mood, which is extracted form Twitter feeds, can affect the Dow Jones Industrial Average (DJIA). They analysed the text content of daily Twitter feeds using two mood tracking tools: The OpinionFinder, which measures negative vs. positive mood, and Google-Profile of Mood States, which measure mood in terms of 6 dimensions (calm, alert, sure, vital, kind, and happy). Their results showed that the prediction accuracy of DJIA is enhanced by the inclusion of specific public mood dimensions. In addition, Sun *et al.* (2016) examined the use of textual data produced from users' micro-blogs in Tweeter to predict the stock market. They were able to find a correlation between the movement of stock prices and the social media content through the usage of the latent space model proposed by Ming *et al.* (2014). Their study did not evaluate sentiment of the social media data, whereas Sorto *et al.* (2017) proposed a sentiment analysis system based on summarisation

to determine the polarity (positive or negative) of news articles from the Wall Street Journal and financial market data from the NASDAQ aimed at predicting the stock market. In addition, Khedr *et al.* (2017) constructed a predictive model to predict stock market future trends. Their model used sentiment analysis of multiple types of financial news and historical stock prices, which led to the achievement of prediction accuracy up to 89.80%. Gálvez and Gravano (2017) mined Argentinian stock message boards to check whether they contain important predictive information. . Hence, they built and validated a set of predictive models using machine learning and topic discovery techniques. The results of their study demonstrated that these predictive information from stock message boards has important semantic content and could enhance the classification performance based on technical indicators. Nardo *et al.* (2016) investigated the influence of online news on the financial market. They conclude that the most promising avenue for research is the metadata of the communication flow, and its properties such as the frequency of posts and their discriminative terms as well as the strength of comments could be in analysing stock markets bubbles.In their forthcoming paper, Baeza-Yates *et al.* (2019) discuss a different application such as entity retrieval and sentiment analysis, in relation to micro-blogs and Twitter messages due to their popularity. Most of the techniques used exploit emoticons and trained on a sample of emoticon based positive and negative tweets.

Relying solely on textual data analysis while studying stock markets has some limitations such as late evaluation for the importance of news events, and the different opinions and interpretations of the events, which the experts may have. In addition, experts and investors may lapse into making relatively less rational decisions in the financial market because of the lack of sufficient and clear information about relationships between decision variables and outcomes. This problem becomes worse when decision makers are confronted with large amounts

of information as in the domain of stock markets analysis (Wang *et al.* 2011). Valuable information for the purpose of investment, which is implicit in textual data is not easy to discern. Hence, despite the wealth of data, many fund managers have been unable to fully capitalise on their value (Kannan *et al.* 2010). The key issue is the necessity to have the user's specification to label historical documents for training and classifying. The use of textual information contains great wealth of knowledge, which is complex and rich. However, it relies heavily on human analysis in order to achieve a better prediction of stock market price movements. Unlike numerical and fixed field data, it cannot be analysed by standard statistical data mining method (Nasukawa & Nagano 2001). Identifying the major events, which have impact on the stock market, and characterising them in order to design strategies for predicting the market is another important problem, which was addressed by Mahajan *et al.* (2008).

2.3 Summary and Conclusions

Countries around the world depend on stock markets for their economic growth. Stock market crashes are unavoidable and are, by nature, preceded by speculative economical bubbles. The increasing importance of stock markets and their direct influence on the economy were the main reasons for deciding to study and analyse stock market crashes, which is the application domain of this research.

The need to determine early warning indicators for both banking and stock market crises has been the focus of study by many economists and politicians. Whilst most research into the identification of these critical indicators applied data mining to uncover hidden knowledge, very few attempted to adopt a text mining approach. Patel *et al.* (2015) explained that stock markets behave randomly; consequently, the application of data mining to the analysis of stock market data may not be sufficient to model and justify any random behaviour of the market.

Given the huge amounts of free news and financial data, it is important to study the rich information embedded in this data, known as “alpha”. This research is an attempt at addressing this issue and discovers the critical indicators from unstructured yet valuable source of information.

Chapter 3 Text Mining Theoretical Basis

3.1 Introduction

In the previous chapter a literature review on the Stock Market (SM) domain applying Data Mining (DM) and Text Mining (TM) techniques to predict its movements was presented. The aim of this chapter is to describe the text mining theoretical basis, which underpins this research and the development of the Stock Market Random Forest-Text Mining (SMRF-TM) approach. The design and implementation of the SMRF-TM approach is presented in chapter four.

The structure of this chapter is as follows. In section 3.2 an introduction of text mining and its main principles are given. In section 3.3 the different stages of text mining are presented; this is followed by an explanation of the different tasks carried out during each stage. Section 3.4 summarises the chapter and presents the conclusions.

3.2 The Main Principles of Text Mining

As discussed in chapter one, text mining research area is considered a subfield of knowledge discovery. It involves the operation of finding interesting, non-trivial and previously unknown rich information (alpha) from different written resources, which are either unstructured or semi-structured text (Drury 2013). The term Alpha represents rich information, which is not defined. In order to be able to retrieve such information, text mining is applied to analyse documents and elicit useful patterns and relationships between its features, in order to discover new knowledge (Drury 2013, Gupta & Lehal 2009, Hearst 2003). Extracting hidden and potentially critical relationships is one of the main advantages of text mining, as this helps users transform large volumes of electronic textual documents into a structured repository of insightful and valuable information. According to Hearst (2003), text mining is

also known as a knowledge discovery approach in text, text data mining or intelligent text analysis.

Knowledge can be retrieved from many different sources of information, but natural language is still the biggest available source as it is usually stored as text (Korde & Mahender 2012, Gupta & Lehal 2009, Hearst 2003). The demand to analyse large volumes of textual data was the main reason for the evolution of text mining and it also gave it a high commercial value. Text mining draws on information retrieval, data mining, machine learning, statistics and computational linguistics (Gupta & Lehal 2009).

There is a great difference between the text mining and data mining approaches. Data mining makes use of the strong internal structure of stored data to extract additional non-trivial useful information. On the other hand, text mining is applied to non-structured or semi-structured text documents since documents rarely have strong internal structure. The need to understand the text, which resides in such documents and extract valuable information for the purpose of analysis required the application of natural language processing techniques in text mining applications. Applying natural language processing techniques on text documents can reduce the size of textual data to a tractable size; this facilitates the analysis of the information contained in these documents to gain new knowledge. Consequently, text mining involves pre-processing of documents, storage of the intermediate representations, techniques to analyse these intermediate representations, and visualisation of results (Miner 2012, Feldman & Sanger 2007, Hearst 2003). In text mining systems, document pre-processing operations are concerned with the identification and extraction of representative features for natural language documents by transforming unstructured data stored in the documents corpus into a more structured intermediate format (metadata). The metadata enriches the content representation of the documents thus supporting mining software to manipulate it.

Text mining can be considered as an extension of data mining (Yu *et al.* 2005). It is a significant process but more complex and challenging than data mining, because natural language is ambiguous, subtle and very rich (Mahgoub *et al.* 2008).

However, text mining still has several drawbacks. First of all, the initial conditions can dramatically affect the final results, for example the way in which the features are identified and represented for further text mining. Second, it requires large human input from the domain experts to validate the system and evaluate the results (Feldman & Sanger 2007, Feldman & Dagan 1995). Finally, results produced may require further refining, as the final solutions (i.e., rules and patterns) may be sometimes uncertain, vague and imprecise (Yu *et al.* 2005).

3.3 The Main Stages of Text Mining

Text mining consists of three main stages, which are information retrieval, information extraction and the analysis of the extracted information using data mining techniques. Information retrieval is to retrieve relevant documents in response to a query and so it is concerned with data collection. Data collection includes gathering, selecting, and filtering of documents, which may prove to be useful for the analysis. In other words, information retrieval targets to find what is already known through (a) specifying a general description of the query, (b) searching the documents collection, and (c) returning subsets of documents relevant to the query (Uppal & Lee 2017, Weiss *et al.* 2010, Feldman & Sanger 2007).

The information extraction stage is responsible for the analysis of textual data, finding relevant entities and discovering facts about these entities through the deployment of Natural Language Processing (NLP) techniques until the desired information is extracted. So, the target of information extraction is to extract important entities for further mining (Uppal & Lee 2017, Feldman & Sanger 2007).

Finally, text mining systems deploy data mining techniques (i.e. decision trees, association rules, clustering, etc.) to find hidden relationships within the extracted features to discover hidden new knowledge. Consequently, the target of using the data mining techniques is to mine the metadata to extract useful knowledge and evaluate the results (Uppal & Lee 2017, Gupta & Lehal 2009, Feldman & Sanger 2007).

Some stages involve a set of tasks. The information extraction stage consists of three tasks: (a) documents pre-processing task including data preparation, noise reduction, tokenisation, stop words removal, stemming and negation expressions handling, (b) features generation task and representation into a vector space, and (c) features extraction task based on Term Frequency/Inverse Document Frequency (TF/DF) and statistical analysis. The text mining stage applies data mining techniques to analyse these extracted features, which become the metadata and evaluates the findings and the discovered new knowledge (Figure 3.1).

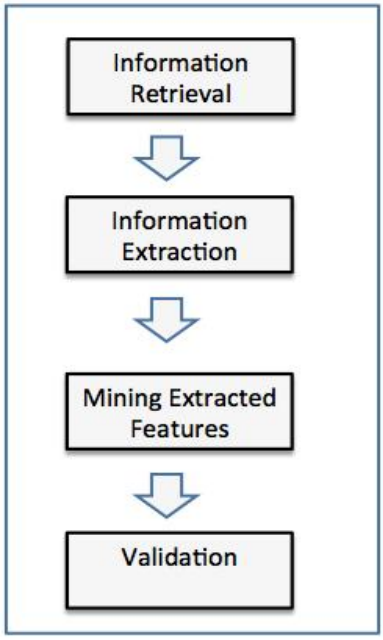

Figure 3.1 The main stages of text mining

3.3.1 Information Retrieval

The first task in text mining is the data collection, which does not depend on limited source set, but, for example, on searching all websites available to search engines. This stage can be done either automatically using a text mining tool or manually by searching the web using special keywords related to the domain of study. First, we need to know the methods available for information retrieval. Generally, retrieval methods handle the retrieval problem of documents either as a selection problem or as a ranking problem (Weiss *et al.* 2010).

The Boolean retrieval model is considered as a common method for document selection where a document is expressed by a set of keywords. The user needs to specify a query in the form of Boolean expression of keywords to retrieve a document. So, the query is presented as determining constraints for selecting relevant documents and the retrieval system would retrieve documents that match the Boolean query expression. The Boolean retrieval system performs well only when the user's knowledge about the document collection is deep, so that s/he can develop a good query in order to determine exactly the user's information needs with a Boolean query (Weiss *et al.* 2010, Drury 2013, Feldman & Sanger 2007).

As for the document ranking methods, documents are ranked according to their relevance depending on the query. According to a user's keyword query information, retrieval systems generate a ranked list of relevant documents. These methods are more convenient than document selection methods for typical users and preparatory queries. There are a variety of ranking methods due to the wide spectrum of mathematical foundations, such as statistics, probability, algebra and logic. While applying any of these methods, the keywords in a query are matched with those in the documents and each document is weighted depending on how well it matches the query. Approximating the percentage of relevance of a document with a weight computed based on information such as the Term Frequency (TF) in

the document and the whole documents corpus is the main target (Weiss *et al.* 2010, Drury 2013, Feldman & Sanger 2007).

3.3.2 Information Extraction

The electronic data for many applications is mostly available in the form of natural language documents instead of structured databases (Gupta & Lehal 2009). Information extraction is responsible for transforming unstructured textual data into a more structured repository in order to be able to analyse it using pattern matching to identify key phrases and relationships within textual data (Gupta & Lehal 2009).

One of the most important issues to be addressed in Information extraction is feature extraction, which involves identifying and extracting key features from textual data so that it can be used as the data and dimensions for analysis. In order to achieve this, feature extraction algorithms may use dictionaries to identify some terms and linguistic patterns (Gupta & Lehal 2009).

Kuntraruk and Pottenger (2001) developed a massively parallel model for feature extraction, which employs unused cycles on networks of PCs/workstations in a highly distributed environment proving that linear speedups in the number of processors are achievable for applications involving reduction operations based on a novel, parallel-pipelined model of execution. However most of the existing key phrase extraction approaches require human-labelled training sets. To address this issue Huang *et al.* (2006) used two novel feature weights, which can be used in both supervised and unsupervised tasks to develop an automatic key phrase extraction algorithm. Their algorithm treats each document as a semantic network, which holds both syntactic and statistical information. By taking advantage of the structural dynamics of these networks they could identify key nodes, which can be used to extract key phrases unsupervised, resulting in 50% improvement in effectiveness and 30% in efficiency in unsupervised and supervised tasks as well.

Liangtu and Xiaoming (2007) presented a novel feature extraction algorithm to improve the efficiency of web texts processing, which is based on the improved particle swarm optimisation with reverse thinking particles. They described the web textual data using vector space model. Wong and Lam (2009) developed an unsupervised learning framework, which can extract information and conduct feature mining over different sites' web pages. It allows tight interactions between the tasks of information extraction and feature mining. They leveraged information from different sources as they simultaneously consider web pages across different sites by using an undirected graphical model, which can model the interdependence between the text fragments within either the same web page or different web pages. A number of supporting tools for feature extraction have been developed. However, they tend to consider text as a simple literal while text is semantically significant and requires a tool, which considers its linguistic characteristics. As a result Myung *et al.* (2009) developed a customised extraction method, which considers the characteristics of source documents called PicAChoo, which stands for 'Pick And Choose'. It provides an environment, which enables feature extraction methods using the structure of sentences and the part-of-speech information of words. They also suggested dynamic composition of different extraction methods without hard coding. In order to enhance machine learning algorithms used in email filtering L'Huillier *et al.* (2010) implemented a feature extraction methodology for phishing emails, which, uses latent semantic analysis features and keyword extraction techniques. They obtained a competitive feature set against previous phishing feature extraction methodologies and they achieved promising results over different benchmark machine learning classification techniques. Feng *et al.* (2011) developed a new keyword extraction algorithm based on sequential patterns, which is independent of languages and does not need to use semantic dictionary to get the semantic features. They did so by presenting a document as sequences of words and applying a sequential pattern-mining algorithm on it and the important

sequential patterns, which reflect the semantic relatedness between words were mined. In order to build their keyword extraction model, they used statistical features as well as pattern features within words. By applying their algorithm on Chinese journal articles, they proved that their algorithm yielded better results than the baseline method keyword extraction algorithm. As mentioned above, information extraction involves many tasks, which are discussed below.

3.3.2.1 Documents Pre-processing

The documents pre-processing task is considered a basic component of any natural language processing system because the words/sentences produced from it are significant constituents passed to all the subsequent text processing stages (Vijayarani *et al.* 2015, Kannan & Gurusamy 2014). Subjecting the text documents to a set of activities in order to eliminate all the words, which are unlikely to support text mining is the goal of documents pre-processing.

Documents pre-processing starts by transforming the raw textual documents into plain text documents by removing all the hash tags, URLs, links and similar elements. This is followed by removing all the undefined characters from the previously generated plain text documents in order to reduce noise within the data corpus. Tokenisation is then applied on these documents, which is a form of text segmentation done by splitting the text streams within these documents into separate words/phrases called tokens. This is achieved by using the white spaces, commas, semi colons, brackets, punctuation marks, exclamation marks, question marks, …etc. to split the text streams into tokens (Vijayarani *et al.* 2015, Kannan & Gurusamy 2014). The list of tokens produced after tokenisation still contains words, which are frequently used and do not contribute to the context of the documents as they are only used to join words/sentences together. Examples of such words are ‘a’, ‘an’, ‘and’, ‘the’; these are referred to as stop words. These stop words should be removed, as they can constitute an obstacle while trying to understand and

classify the documents. Stemming follows stop words removal, which is the process of transforming all the different forms of a word into their root format (stem). Last but not least, there is the handling of negation expressions.

3.3.2.2 Features Generation

The features generation task is responsible for generating a 2D vector space containing the entire root formats of the words, which are left after text pre-processing. This vector space is considered the basic input for the next text mining task where the columns represent the features, the rows represent the documents and the cells contain the root formats of the words in the documents.

3.3.2.3 Features Extraction Using Term Frequency and Inverse Document Frequency

Revealing hidden information and relations in text is the main target of features extraction. A semantic analysis based approach or text-weighting approach can be used to solve text features extraction problems. Features extraction normally follows features generation in order to exclude features, which do not provide valuable information. So, features extraction yields a reduced dimensional vector space representation (Feldman & Sanger 2007).

Among the most popular indicators are Term Frequency (TF), Inverse Document Frequency (IDF), and their multiplicative combination (TF-IDF). In the TF approach, the assumption is that the words occurring more often in a document are more important than other words. In the IDF approach, the biggest explanatory power is believed to exist in the rarest words in the document collection. The two measures are combined into TF-IDF, which is generally considered a basic indicator used in features extraction (Hakim *et al.* 2014, Chakraborty 2013). To achieve good performance Wei and Dong (2001) suggested that, at the end of the selection process, only words with the highest TF-IDF score should be selected as features.

The vector-space model can be used to represent a document. A document can be represented as a vector *(v)* in the *(t)* dimensional space given a set of *(d)* documents and a set of *(t)* terms. Since TF is the number of occurrences of term *(t)* in the document *(d)*, which is denoted as *freq* (*d*,*t*). Then the weighted TF matrix *TF* (*d*,*t*) measures the association of a term *(t)* with respect to document *(d)*: it is usually set to (0) if the term does not exist in the document, and set to the TF *freq* (*d*,*t*) otherwise. The IDF is another relevant measure, which represents the importance of a term *(t)*, where the importance of a term *(t)* will be reduced if it occurs in many documents due to its low discriminative power (Hakim *et al.* 2014, Chakraborty 2013).

Regarding the mathematical formulation of IDF, Croft and Harper (1979) formulated an equation for IDF based on the binary independence model. Effectiveness of IDF was theoretically validated by Greiff (1998) through arguing the relationship between pair-wise mutual information and IDF. One year later, Church and Gale (1999) showed that larger IDF values mean larger deviations from the Poisson distribution and so more 'context' regarding the terms. In order to do so they tested the gap between observed and predicted IDF values using empirical studies.

Consequently, to enhance the performance TF and IDF are combined forming the TF-IDF measure in a complete vector-space model as shown below:

$$TF\text{-}IDF\,(d,\,t) = TF\,(d,\,t) \times IDF(t)$$

3.3.3 Analysis of Extracted Features Using Data Mining Techniques

The vector space model constructed by an information extraction stage is provided to the data mining stage where its techniques are applied to mine the extracted information, discover new implicit knowledge and derive new facts. The

most common data mining techniques used to mine the vector space model produced by the information extraction stage are mining the metadata to extract useful knowledge, the analysis and the evaluation of the discovered knowledge (Uppal & Lee 2017, Witten *et al.* 2016, Gupta & Lehal 2009, Feldman & Sanger 2007).

There are two data mining techniques used for machine learning: supervised learning or unsupervised learning. Document classification is an example of supervised machine learning in the form of natural language processing, where a model is created based on a training set. Categories are predefined and documents within the training dataset are manually tagged with one or more category labels. A classifier is then trained on the dataset, which means it can predict the category of a new document. The aim of classification of documents is to assign one or more classes to a document, which makes it easier to manage and sort the documents within the data corpus (Jabeen *et al.* 2018, Witten *et al.* 2016, Ghaffari *et al.* 2013, Miner 2012, Kamruzzaman *et al.* 2010). Text classification is the process of classifying documents into predefined classes based on their content. It is the automated assignment of natural language texts to predefined classes. Existing supervised learning algorithms applied to automatically classify text need sufficient numbers of documents to learn accurately. The words within a document (features) can be used to support the prediction of the classification of a new document (Jabeen *et al.* 2018, Ghaffari *et al.* 2013, Kamruzzaman *et al.* 2010).

Unsupervised learning is the other technique for machine learning, which is used to draw presumptions from datasets containing unlabelled data. Unlike supervised learning algorithms there is no evaluation of the accuracy of the output of the unsupervised learning algorithms, since the data given to the learner is unlabelled. The most common unsupervised learning method is cluster analysis. Clustering is used either for exploratory data analysis to find hidden patterns or for data grouping (Witten *et al.* 2016).

3.3.3.1 Supervised Classification

Supervised classification, which is applied in our SMRF-TM approach, involves dividing the records into predefined categories. There are multiple methods, which are popular such as Bayes, rules and trees classifiers.

- **Bayes classifiers**

The naïve Bayes method is a kind of module classifier under known priori probability and class conditional probability (Korde & Mahender 2012). The basic idea behind naïve Bayes is to calculate the probability that document D belongs to class C. There are two event models for naïve Bayes, which are the multivariate Bernoulli model and multinomial model (Vidhya & Aghila 2010, McCallum & Nigam 1998, Lewis 1998). Out of these models, the multinomial model is more suitable when the database used is large but there are two serious problems with the multinomial model. The first problem is rough parameter estimation and the difficulty of handling rare categories, which contain only few training documents. Kim *et al.* (2006) proposed a Poisson model for naïve Bayes text classification and they also used a weight enhancing method to improve the performance of rare categories. Modified naïve Bayes is proposed by (Shen & Jiang 2003) to improve performance of text classification. Naïve Bayes is easy to implement and compute but its performance is very poor when features are highly correlated and it is sensitive to features selection (Korde & Mahender 2012).

- **Rules classifiers**

The decision rules classification method uses rule-based inference to classify documents to their annotated categories (Korde & Mahender 2012, Apte *et al.* 1994). A popular format for interpretable solutions is the disjunctive normal form model. In the case of handling a dataset with large number of features for each category, heuristics implementation is recommended to reduce the size of rules set without affecting the performance of the classification. Wu (2009) presented a

hybrid method of rule based processing and back-propagation neural networks for spam filtering (Korde & Mahender 2012).

- **Tree classifiers: namely decision trees and random forest**

A **Decision Tree (DT)** consists of tree internal nodes, which are labelled by term, branches departing from them labelled by a test measure on the weight, and leaf nodes representing corresponding class labels. A decision tree can classify a document by running through the query structure from root until it reaches a certain leaf, which represents the goal for the classification of the document. Most of the training data will not fit in memory decision tree construction as it becomes inefficient due to swapping of training tuples. To handle this issue Mehta and Agrwal (1996) presented a method, which can handle numeric and categorical data. Johnson *et al.* (2002) presented decision-tree-based symbolic rule induction system for text categorisation, which improved text classification. The decision tree classification method is used in many applications because it has a number of advantages over other decision support tools, such as greater simplicity in understanding and interpretation even for non-expert users (Patel 2017, Korde & Mahender 2012, Chen *et al.* 2010).

Random Forest (RF) classifier is another tree classifier, which could be used to classify text and it is often considered and applied as embedded features selection method in text mining, because of the ability to measure descriptor importance as well as similarity between features. RF combines the bagging approach with a random sub-sampling method and so it is treated as a special modification of bagging. While bagging works with any algorithm as a weak learner, random forest is an ensemble classifier consisting of many decisions trees and output the mode of the classes' results by individual trees. Similar to bagging, RF is easily comprehensible and can reduce the variance of the prediction accuracy, but due to the sampling of attributes, the learning process of random forest is usually faster. Random forest can handle a very large number of input variables, and even when a

large portion of attribute values is missing, it is often able to maintain the desired accuracy. Just like bagging, random forest selects instances randomly with replacement (bootstrap), but unlike bagging, random forest samples attributes without replacement for each tree. The trees are grown to maximal depth without pruning and each tree performs an independent classification/regression. Each tree is then assigned a vector of attributes or features to a class and the forest chooses a class having the most votes over all trees using a majority vote or averaging. The commonly used growing algorithm for the single decision tree is the Classification and Regression Tree (CART) (Romo & Araujo 2013). Each tree is grown as follows: If the number of cases in the training set is N, sample N cases at random with replacement (i.e., the size of the sample is equal to the size of the training set but some instances of the training set may be missing in the sample while some other instances may appear more than once in the sample). This sample is the training set for growing the tree. If there are M input variables, a number m < M is specified such that at each node m variables are selected at random out of the M and the best split on these m attributes is used to split the node. The value of m is held constant during the forest growing. Each tree is grown to the largest extent possible without pruning.

Breiman (2001) has shown that random forest can be used for measuring the importance of features, as it is similar to the Adaboost algorithm, which is also an ensemble technique but uses a different approach. It uses classification trees as its sub-component rather than iteratively training near examples previously missed by the classifier. So, the trees are grown using bootstrapped versions of the data and by choosing k nodes for which to search for a split. This introduces random perturbations into the data, which generate different results in each tree and prevents over-fitting (stewart & Zhukov 2009). Knowing that the increase in error due to perturbing feature values in a data set and then processing the data through

the random forest is an effective measure of the relevance of a feature (Cunningham 2007).

Various studies on random forest by Svetnik *et al.* (2004), Dietterich (2002) and Breiman (2001) have shown that the performance of decision trees could be improved if ensembles of trees were used. Svetnik *et al.* (2003) have combined random forest with a feature selection algorithm based on measuring the importance of single features, and successfully applying this combination to the task of QSAR-modelling. Prinzie and Poel (2008) used random forest for multi-class classification and regression by combining it with multi-nominal logit, which is a generalisation of logistic regression, which allows more than two discrete outcomes and is commonly applied within the customer relationship management domain. Biau *et al.* (2008) discussed several consistency theorems for various versions of RF and other randomised ensemble classifiers. Then Arevalillo and Navarro (2011) used random forest to uncover bivariate interactions in high dimensional small datasets (Janecek 2009). Zhao *et al.* (2012) showed that random forest classifier has the best performance at all time.

Even though the predictions of random forest have the drawback that they are the outcome of a black box, especially if a small number of informative variables are hidden among a great number of noisy variables. Random forest is prone to over fitting if the data is noisy, and the CART algorithm used for growing the single trees within random forest does not handle large numbers of irrelevant attributes as well as decision tree algorithms, which use entropy-reducing splitting criteria (Janecek 2009).

Random forest has proven to be very effective when deployed by Hillenmeyer *et al.* (2010) to develop an algorithm predicting protein targets of chemical compounds. As they gathered two training sets, one expert created a set of 83 yeast protein-compound interactions and another yeast homologous of 180 human drug-protein pairs defined as interacting in DrugBank in order to produce random negative

interaction sets. They produced these sets in two ways, balanced (number of negative and positive interactions are equal) and unbalanced (contains all negative and positive interactions). Then they used the Weka software and applied various machine learning algorithms (Bayesian Network, Naïve Bayes, Decision Tree Decision Stump, Logistic Regression, Support Vector Machine and random forest) using 10-fold cross validation on the produced drug-target interactions and features from the balanced and unbalanced sets. The random forest algorithm resulted in the best performance among the machine learning algorithms tested when they trained their algorithm on a segment of the known drug-target interactions, and tested it on another defined segment of the known drug- target interactions. Also, Percha *et al.* (2012) demonstrated the strength of random forest when they used it while applying text mining to predict new Drug-Drug Interactions from the identified gene-drug relationships. They deployed the random forest to classify all the drug pairs in their training set and it surpassed both the SVM and logistic regression classifiers. They used the Out-of-bag estimation of the error to evaluate the performance of the random forest in their training data and they found that random forest provided a natural measure for its classification. By evaluating the paths for a certain drug pair depending on the number of “yes” votes, which each got from the random forest, they were able to decide which paths describe the mechanisms of interaction for that pair. So as the random forest was trained by a set of known interacting drug pairs it could be applied to a new dataset (doesn’t contain any drugs from the training set) and be able to predict if any other pair of drugs will interact or not. This provided them with an efficient way to predict the mechanisms of interaction, which were not yet known proving that their random forest classifier can explain known Drug-Drug Interactions and discover novel Drug-Drug Interactions which were not yet been discovered.

Recently, to predict future values of two SM indices (CNX Nifty and S&P Bombay stock exchange) Patel *et al.* (2015) introduced a two-stage fusion of machine

learning techniques. The first stage uses Support Vector Regression (SVR) while the second stage uses SVR, Artificial Neural Network (ANN) and RF yielding SVR-SVR, SVR- ANN and SVR-RF fusion predictions models. These two stages fusion predictions models were proposed in order to enhance the stock market prediction performance by bridging the gap (g) existing in the available SM prediction methods as they use the statistical parameters' value of $(t)^{th}$ day to predict the $(t+g)^{th}$ day's closing value where the performance decreases as (g) increases. Their experimental results proved that the two stages hybrid model yielded better performance especially in the cases of SVR- ANN and SVR-RF the enhancement was very significant. Hence, they recommended for future research to apply these techniques on textual data (news) since the news about investors' interests, companies' performances and government policies can dramatically affect stock market price movements.

Nikfarjam *et al.* (2010) stated that there are technical analyst researchers who claim that historical market movements tend to repeat themselves and there are visual patterns in a market graph, which can be detected using mathematical models and pattern recognition techniques. However, they tend to state that patterns exist and do not interpret these patterns. They use techniques such as the moving average rules, relative strength rules, filter rules and the trading range breakout rules (Nikfarjam *et al.* 2010). Yu *et al.* (2013) showed in their study that these rules failed in their predictive power. Other more sophisticated financial prediction techniques based on machine learning algorithms such as neural networks (Sermpinis *et al.* 2013, Ghazali *et al.* 2011, Vanstone & Finnie 2010, Anastasakis & Mort 2009), fuzzy logic (Aladag *et al.* 2014, Bahrepour *et al.* 2011), Support Vector regression (Nassirtoussi *et al.* 2015, Premanode *et al.* 2013) and rule-based genetic network programming (Mabu *et al.* 2013) have shown better results. Hillenmeyer *et al.* (2010) have achieved promising results when they have

applied random forest to predict protein targets of chemical compounds. In addition, Ali *et al.* (2012) have compared the classification results of the random forest and the J48 decision tree by applying both of them on the breast cancer data set. The comparison's results showed that the percentage of correctly classified instances for random forest increased from 69.23% to 96.13% as the number of instances increased from 286 to 699, proving that the random forest outperforms the J48 when the number of instances increases (large datasets). Consequently, the random forest is applicable for large datasets modelling as it can deal with missing values and all kind of data (categorical, binary and continuous).

Random forest is efficient, interpretable and achieves accurate predictions for various types of datasets because it uses ensemble strategies and random sampling. The model interpretability and the prediction accuracy of random forest are very rare among most of the machine learning algorithms. Furthermore, random forest is less responsive to outlier data in training data and there is no need to prune the trees because the bootstrapping and ensemble scheme makes random forest capable of overcoming the problems of over fitting. So, random forest has all the advantages of decision trees but it achieves better results most of the times due to its utilisations of bagging on samples, random subsets of variables and voting schemes (Kumar & Khatri 2017, Horning 2013, Qi 2011, Breiman 2001). These have motivated this research to adopt random forest for mining the metadata to investigate its effectiveness in analysing articles related to the stock market crisis, to identify the critical indicators and to label them.

3.3.3.2 Unsupervised Classification

In this section, an overview of unsupervised classification is presented. Unsupervised classification is also known as clustering. The goal of clustering is to distribute a set of data records into groups having high similarity. In text mining, clustering techniques are often dissecting data rather than clustering it through

generating homogeneous areas of data instead of finding existing clusters (Xu & Wunsch 2005, Lebart 2004).

Clusters are usually described by studying their internal consistency and their difference to the other groups (Xu & Wunsch 2005). Clustering can be explained through four steps:

1. Feature selection or extraction: The selected features help in determining the specific patterns the differ one cluster from another
2. Clustering algorithm: The selected features are combined according to certain criteria
3. Cluster validation: Evaluation criteria are applied to the selected clusters thus providing the user with a certain measure of confidence. The criteria used should be neutral and irrelevant to the clustering algorithm used.
4. Results interpretation: Experts interpret the resulting clusters and provide to the user meaning behind the choice according to the original data.

Clustering process includes a feedback where sometimes re-grouping and re-evaluation is applied A broad distribution of the clustering techniques, according to the way clusters are generated, is hierarchical and partitional. Hierarchical clustering adds the features to predefined clusters while partitional clustering splits the features into a predefined number of clusters without a specific structure (Xu & Wunsch 2005).

Clustering algorithms differ from one another according to the starting points used to building the clusters and the criteria according to which the features are divided (Xu & Wunsch 2005). There are many clustering algorithms, which could be applied to the dataset. Two of the most popular types are the K-means and the Expectation Maximisation.

i)K-means clustering

The idea behind this algorithm is to output k-clusters while fulfilling the criteria of minimising the squared-error. The algorithm begins by randomly choosing k objects as the centres of the k clusters. Afterwards, the algorithm repeats the assigning process of each object to the closest cluster while using the mean value of the other objects in the cluster. This results in updating the mean value of the objects existing in each cluster. This process is repeated until there is no resulting changes appear in the clusters (Jung *et al.* 2014).

ii) Expectation maximisation

Expectation Maximisation (EM) is a sub-area of the Gaussian mixture model, which aims to improve the density of the chosen clusters. The EM algorithm starts by specifying the number of clusters and the stopping tolerance. The output is k-clusters having a maximum log-likelihood among its weights. It works in two steps, first the expectation step where the membership probability of every object with each cluster is calculated. After the expectation step, the maximisation step is applied which updates the mixture model parameter. The two steps are repeated until the stopping criterion is fulfilled (Jung *et al.* 2014).

3.4 Validation

Since textual data is mostly large, high dimensional, categorical, and sparse, it produces a huge amount of metadata. Validation procedures in such a complex environment are relatively difficult to be applied, but yet they are still very important. External validation is the most common used procedure in the case of supervised learning models for classification. External validation is usually applied using cross-validation methods to estimate the parameters of the model in the learning phase and to assess the model in the generalisation phase. External validation can be used in the unsupervised models as well but only in two cases: (a) the data set is

big enough to split into subsets, so that subset(s) can be used to learn the model and the other subset(s) can be used to validate the model, (b) the availability of enough metadata or external information to complement the description of the features to be analysed (Lebart 2004).

3.5 Summary and Conclusions

Text mining, which is a subfield of knowledge discovery, involves the pre-processing of document collections, the extraction and representation of relevant features, the application of appropriate data mining techniques to analyse these intermediate representations through the application of supervised/unsupervised algorithms on these representations to discover new knowledge.

Textual data limits the success in the investigation of stock markets because natural language is ambiguous, subtle and very rich. However, this thesis claims that mining stock markets news can enhance the performance of current trading systems' strategies as rich and valuable information is embedded in financial news and need to be discovered.

The literature review showed that the random forest classifier, which is a supervised learning approach, has a number of strengths; this makes it worthwhile to further investigate and apply it to analyse stock markets articles. Random forest can be a good predictor of stock markets because it uses ensemble strategies and random sampling. It is also less responsive to outlier data in training data and the bootstrapping and ensemble scheme help random forest overcoming over fitting. These features have motivated this research to adopt random forest and investigate its effectiveness in identifying critical indicators and evaluating their semantic contribution to the stock market movements. Consequently, in SMRF-TM approach proposed in this research we used supervised classification through the application

of random forest to classify the extracted features as well as the news articles into predefined categories.

Clustering, which is an unsupervised learning algorithm, and the application of expectation maximisation, in particular, is relevant to our text mining stage as it can be used to distribute a set of data records into clusters sharing high similarity, by calculating their membership probabilities. Consequently, we used unsupervised classification through the application of expectation maximisation clustering technique to distribute a set of data records into groups having high similarity. In SMRF-TM approach, the expectation maximisation clustering technique is applied after classifying the extracted features (critical down, down, neutral, up, critical up) using random forest. This is done to cluster the classified features according to their semantic meanings (economic, social, political), which supported SMRF-TM approach to describe the causes behind the classification of the features. In addition, we applied the k-folds cross validation technique in order to evaluate the learning capabilities of SMRF-TM approach, which is a semi-supervised system.

In this research, we develop a semi-supervised system to extend the classification of stock market financial news articles into five meaningful and semantic classes: critical down, down, neutral, up and critical up. This may significantly enhance the prediction performance because financial news contains valuable statistical parameters, which can be very useful for the purpose of analysis. The design and implementation of the Stock Market Random Forest-Text Mining (SMRF-TM) approach adhere to the above described text mining stages, and are presented in the following chapter.

Chapter 4 Stock Market Random Forest-Text Mining (SMRF-TM)

4.1 Introduction

The aims of this chapter are to describe the architecture of the proposed Stock Market Random Forest–Text Mining (SMRF-TM) approach based on the findings of the literature review and the text mining theoretical basis, to describe the three development stages of SMRF-TM approach, explain the two phases of the implementation, and to analyse the results yielded by each phase. The SMRF-TM approach is designed and implemented based on the three text mining stages (Figure 3.1) and captured in Figure 4.1.

The structure of this chapter is as follows. Section 4.2 describes the three stages, which were adopted for developing the SMRF-TM architecture. Section 4.3 discusses the Information Extraction (IE) stage and explains the different tasks carried out in this stage. Section 4.4 describes the semantic analysis of the extracted features, which is executed on (***VS4)*** yielded from the second stage of the research in order to classify and cluster semantically the extracted features and their corresponding news articles. Then, in section 4.5, the two phases of the implementation for the SMRF-TM approach and the results yielded are discussed and the analysis of these results is included. The final section in this chapter is section 4.6, which summarises the main findings.

4.2 SMRF-TM Development Stages

In stage one, which is the Information Retrieval (IR) stage, the textual data selected to test our approach was obtained through a formal subscription in the official web site of the Financial Times. A total of 544 financial news articles

concerning Dubai's stock market, published in the period between 2008 till 2012, were retrieved. This specific period was chosen so that it includes articles published before the crisis and after the crisis within the period for recovery of Dubai's stock market (Dubai's SM upturn). These 544 articles, which have around 1031006 total number of words are used for training and testing and served the basis to investigate the validity of the proposed SMRF-TM approach.

Stage two, which is the Information Extraction (IE) stage starts by performing text pre-processing through the deployment of natural language processing tasks and it is implemented using MATLAB 2011, followed by features generation. The last task to be performed in stage two is the Feature Extraction (FE); this begins by computing Term Frequency (TF), Inverse Document Frequency (IDF) and Term Frequency/Inverse Document Frequency (TF/IDF). Term frequency is the number of occurrences of term *(t)* in the document *(d)* and is represented in the following form *freq* (*d*,*t*). The term frequency is placed into the weighted term frequency matrix *TF* (*d*,*t*) thus showing the association of a term *(t)* with a given document *(d)*. Each cell in the weighted matrix is set to (0) if the corresponding term does not exist in the document, and to the term frequency *freq* (*d*,*t*) otherwise. The Inverse Document Frequency (IDF) is another relevant measure, which represents the importance of a term *(t)*, where the importance of a term *(t)* will be reduced if it occurs in many documents due to its low discriminative power. Regarding the mathematical formulation of IDF, as explained previously in chapter three that Croft and Harper (1979) formulated an equation for IDF based on the binary independence model. Then the effectiveness of IDF was theoretically validated by Greiff (1998) through arguing the relationship between pair-wise mutual information and IDF. One year later, Church and Gale (1999) showed that larger IDF values mean larger deviations from Poisson's distribution and so more 'context' regarding the terms. In order to do so they tested the gap between observed and predicted IDF values using empirical

studies. Consequently, to enhance performance TF and IDF are combined together forming the TF/IDF measure in a complete vector-space model as shown below:

$$TF/IDF\ (d, t) = TF(d, t) \times IDF(t)$$

TF/IDF is a term weighting matrix, which is broadly used in today's information systems.

Stage three focuses on semantic analysis of these extracted features to reveal hidden knowledge and relations between these extracted features, supported by applying Random Forest (RF) classifier and Expectation Maximisation (EM) clustering technique. One of the main novelties of the SMRF-TM approach is the application of random forest classifier on the domain of stock market textual data. This stage classifies the extracted features and the news articles using a set of classifiers, namely RF, ADTree, J48, J48graft, Decision Stump, Random Tree, Bayes Net, Bagging, Rotation Forest, Decision Tables followed by clustering them using the expectation maximisation clustering technique, which are supported by the software WEKA. The use of a set of different classifiers provides a comparative study between the results yielded by other type of classifiers and the results produced by the random forest.

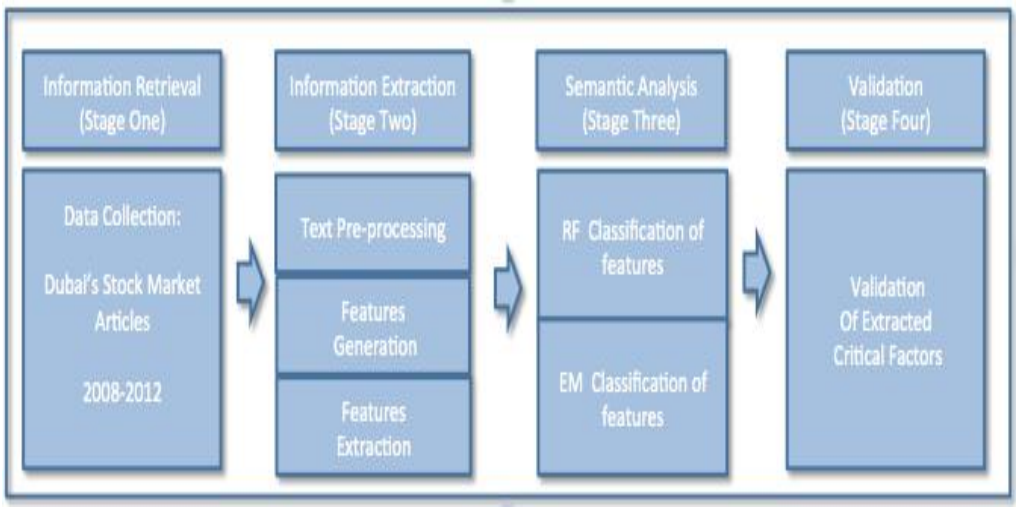

Figure 4.1 SMRF-TM Architecture

4.3 Information Extraction (Stage two)

Stage two of the SMRF-TM approach architecture, which is the information extraction stage, focuses on processing the textual data retrieved in stage one using natural language processing analysis to extract the relevant features, which describe best the movements of the stock markets. This involves a number of pre-processing tasks but the most commonly used are data preparation (transformation into plain text and noise reduction), tokenisation (representing the documents in unigrams/bigrams words), stop words removal, stemming and negation words handling. Unigrams are N-Grams of size one (single word) and the bigrams are N-Grams of size two (two words). Bag of words is another notation used for unigrams/bigrams features (Wang *et al.* 2011). Features generation and features extraction follows the text processing tasks. Tasks performed in stage two are shown in Figure 4.2 and Figure 4.4.

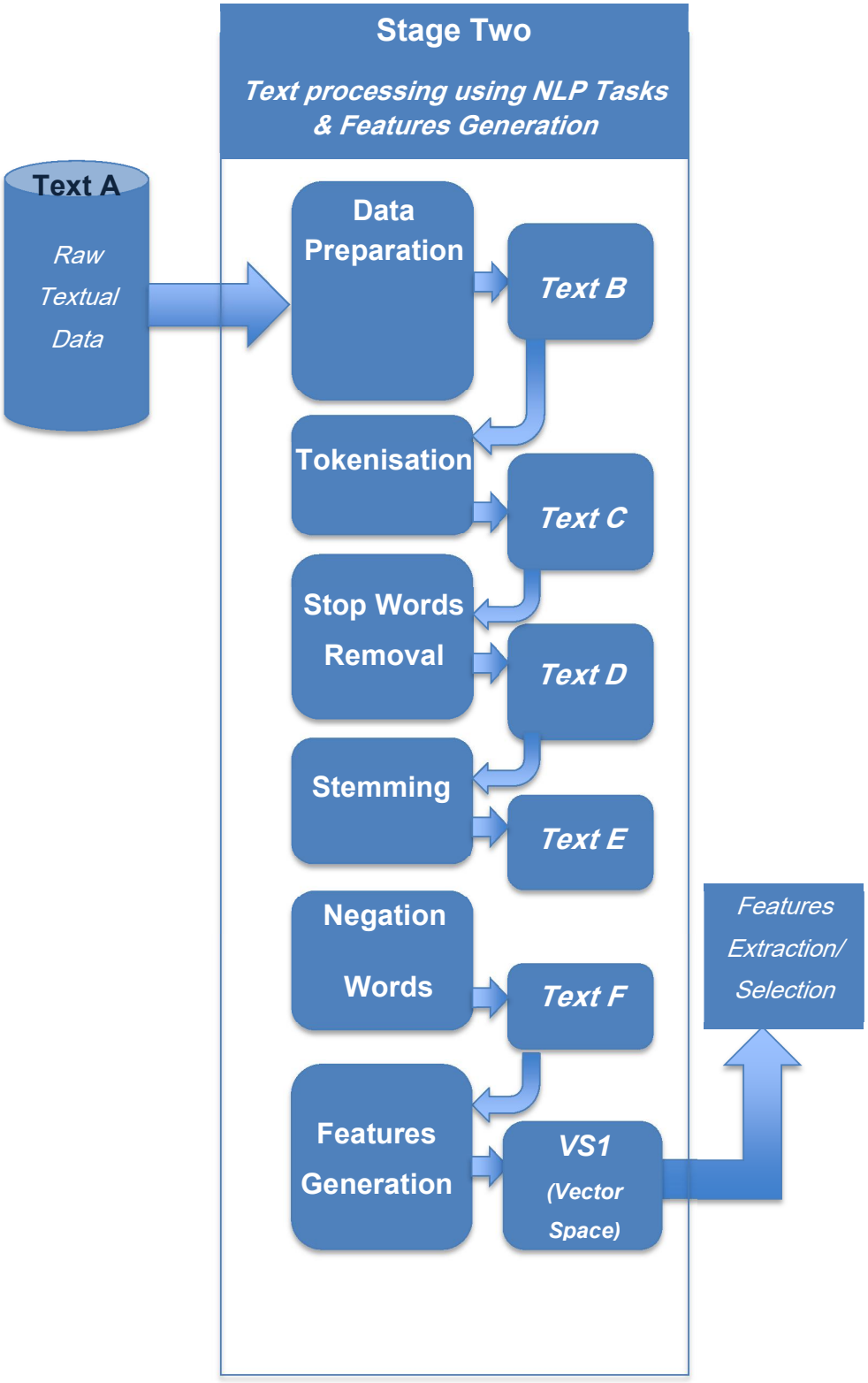

Figure 4.2 Text Processing Stage

4.3.1 Data Preparation

The collected raw data was in web format text. Consequently, it needed some preparation before applying text mining tasks. First all the hash tags, URLs and links were removed. Then the format of the raw data had to be transformed into plain text, which revealed some undesired noise in the text such as undefined characters. So, this process had to be followed by a noise reduction process in order to delete all these irrelevant undefined characters producing ***Text B***, which is sent to the next documents pre-processing task (tokenisation).

4.3.2 Tokenisation

The process of splitting a text stream into words or phrases (tokens) is called tokenisation. Tokens are also considered as input for features extraction processes.

Text streams have many ways to be tokenised. The simplest way is to split the text on blank spaces, but in this research the punctuation and other signs such as ('.', ',', '-', ';', ':', '"', ')', '(', '?', '@', '+', '&', '[', ']', '*', '<', '>', '\', '/', '{', '}', '~' and '!' are also used as they do not hold any significant information for the purpose of semantic analysis. The output of the tokenisation task is ***Text C***, which is sent to the next text processing task (stop words removal).

The implementation of SMRF-TM is carried out in two phases:

(i) In phase one, tokenisation produced 13,061 unigrams tokens extracted from the initial 161 news articles.

(ii) Phase two, which expanded the textual data to 544 news articles, consisted of further two experiments. Experiment one produced 15,276 unigrams tokens while experiment two, which focused on bigrams tokens, yielded 103,506 tokens.

These unigrams and bigrams tokens are stored in ***Text C*** as a one-dimensional array.

4.3.3 Stop Words Removal

Words, which occur frequently but do not carry any significant information, are called stop words; these include determiners, prepositions, pronouns, auxiliary verbs and conjunctions, etc. They are removed to reduce the amount of noise and strengthen the number of relevant features. The application of stop words removal in phase two has reduced the 13,061 unigrams tokens to 12,790 tokens, and in phase two, experiment one, from 15,276 to 15,002 tokens. In experiment two, it has reduced the bigrams tokens from 103,506 to 87,260 tokens. The output of the stop words removal task is stored in ***Text D***, which is sent to the next stemming task.

To remove stop words, researchers may create a domain dependent stop words list by removing high and low frequency words, or by using any statistical measure like information gain, chi-square or TF/IDF. In SMRF-TM the stop words list is created and TF/IDF statistical measure is applied.

4.3.4 Stemming

Stemming is the contraction of a word from its altered form to its root or basic form. The stem is not always the linguistic root of the words. But the most important thing is that related words map to the same stem, even if this stem is not an existing linguistic root. Table 4.1 below shows some examples of such stems, which are not an existing linguistic root in English language, in addition to their original interpretations within the applied textual dataset.

In some cases, stemming might reduce the efficiency of a text classifier. However, many researchers state that although stemming reduces the dimensionality of features and makes the data less sparse and faster to work with, it can enhance the effectiveness of a text classifier (Baker & McCallum 1998). In the implementation of the SMRF-TM architecture we applied the Porter stemmer, which uses a set of language specific rules to transform a word into its basic form. In

phase one the tokens count was reduced from 12,790 to 8,770 tokens, in experiment one of phase two the tokens were reduced from 15,002 to 10,501 unigrams tokens and in experiment two of phase two the tokens were reduced from 87,260 to 82,814 bigrams tokens. The output of the stemming task is stored in ***Text E***, which is sent to the next documents pre-processing task (negation words).

Table 4.1 Examples of stemmed words, which are not an existing linguistic root in English language and their original interpretations

Stemmed Words	Original Interpretations of the Stemmed Words
manag	management, managements, manager, managers, managing, manageable and managed
servic	service, services, servicer, servicers and servicing
inflat	inflation, inflations and inflationary
altern	alternative, alternatives and alternating
financi	financial, financing, refinancing and financier
industri	industries, industrial and industrialised
privat	Private, privatise and privatisation
practic	Practice, practical and practicing
experi	experience, experiences and experienced
princ	principle, principles, prince and principal
opportun	opportunity and opportunities
equiti	equities
crisi	crisis

4.3.5 Negation Words

Negation words within textual data are another very important issue to be considered in text processing tasks. Some of the common negation words such as no, not, n't, neither and nor are removed and the word (not) is concatenated to its following term, for example, "don't increase" becomes "notincrease", "no interest" changes to "notinterest", "not good" is replaced by "notgood", …etc. This task was added based on the expert's recommendation at a later stage. The experimental works were done once without handling negation words and another time after modifying the code to handle negation words. Consequently, the tokens count has increased from 8,770 to 9,198 unigram tokens in phase one, from 10,501 to 11,036 unigram tokens in phase two experiment one, but the count decreased from 82,995 to 82,814 bigrams tokens in phase two experiment two as a result of the merged tokens. Applying RF classifier to these tokens the classification accuracy is enhanced in the three experiments: it increased from 84% to 88.82% in phase one, from 92.28% to 98.35% in phase two experiment one and from 89.71% to 98.89% in phase two experiment two. Table 4.2 summarises the number of tokens produced at each task for each phase.

Table 4.2 Number of tokens produced at each task for each phase

Tasks	Tokens Count		
	Phase 1	Phase 2 Experiment 1	Phase 2 Experiment 2
Tokenisation	13,061	15,276	103,506
Stop Words Removal	12,790	15,002	87,260
Stemming	8,770	10,501	82,995
Negation Words Handling	9,198	11,036	82,814
Features Generation	9,198	11,036	82,814
Features Extraction	709	1057	5987

4.3.6 Features Generation

This task is applied on ***Text F***, which is generated by the task of negation words handling. The goal of the features generation is to transform ***Text F*** into a vector space model representation ***VS1*** for further analysis. The vector space ***VS1*** serves as the basic input for the features extraction where the rows represent the documents and the columns represent the features' root formats in the documents as shown below in Figure 4.3.

		Features			
Documents	D_1	F_{11}	F_{12}	.....	F_{1x}
	D_2	F_{21}	F_{22}	.....	F_{2y}
	⋮	⋮	⋮	⋮	⋮
	D_n	F_{n1}	F_{n2}	.....	F_{nz}

Figure 4.3 The shape of vector space VS1

4.3.7 Features Extraction

This section explains the features extraction techniques used to select the most appropriate features representing the stock market articles for further analysis and mining. A text-weighting approach or semantic analysis based approach can be used to solve text features extraction problems. Features extraction reduces high-dimensionality by only selecting the most useful features. Extracted textual features can be unigrams, bigrams, noun phrases, proper nouns or name entities. SMRF-TM focuses on the unigrams and bigrams features. Among the most commonly used features extraction matrices are information gain, mutual information, odds ratio, correlation coefficient, chi-square and Term Frequency/Inverse Document Frequency (TF/IDF) (Taşcı & Güngör 2013, Taşcı & Güngör 2008, Forman 2007,

Forman 2003). The TF/IDF is applied in stage two of SMRF-TM as shown below in Figure 4.4. Regarding the Term Frequency (TF), the increase in the TF of a word in a specific document indicates its importance. On the other hand, the biggest explanatory power in the Inverse Document Frequency (IDF) exists in the rarest words in the document collection. The features extraction process is dependent on the combination of those two measures into TF/IDF, which is the multiplicative combination of TF and IDF (Yu *et al.* 2005). For good performance Wei and Dong (2001) suggested that, at the end of the selection process, only words with the highest TF/IDF score are selected as features.

A vector-space model is used to capture the relevant extracted features for each article/document within our data. We can represent each document as a vector *(v)* in the *(t)* dimensional space if we have a set of *(d)* documents (i.e. articles) and a set of *(t)* terms. The features extraction stage produces a two-dimensional vector space where the rows represented the articles and the columns represented the features, and the cells capture the TF/IDF value for each feature. In phase one, the vector space captured 161 articles and 9,198 features. In phase two experiment one, the rows represented the 544 articles and the columns represented 11,036 unigrams features, whereas, in phase two experiment two the rows represented the 544 articles and the columns represented 82,814 bigrams features.

SMRF-TM applies TF/IDF to remove all the tokens with a threshold less than a set of different values and the results yielded from all these values were compared to check for the best threshold to be set. As shown below in Tables 4.3, 4.4 and 4.5, the best classification accuracy of the RF is 88.82% in phase one of the implementation; this is produced by setting a threshold to >2, which reduced the number of the features from 9,198 to 709 features. In addition, the best classification accuracy of the RF 98.34% in phase two experiment one is achieved by setting a threshold to >2 as shown below in Tables 4.6, 4.7, 4.8 and 4.9, which

reduced the number of features from 15,276 to 1,056 features. On the other hand, in phase two experiment two the best classification accuracy of the RF is 98.89% by setting a threshold to >1 as shown below in Tables 4.10, 4.11 and 4.12, which reduced the number of features from 82,814 to 5,988 features. The results show that regarding the bigrams tokens the SMRF-TM only performs better when the number of the extracted features increases and the performance is dramatically affected when the number of the extracted features decreases.

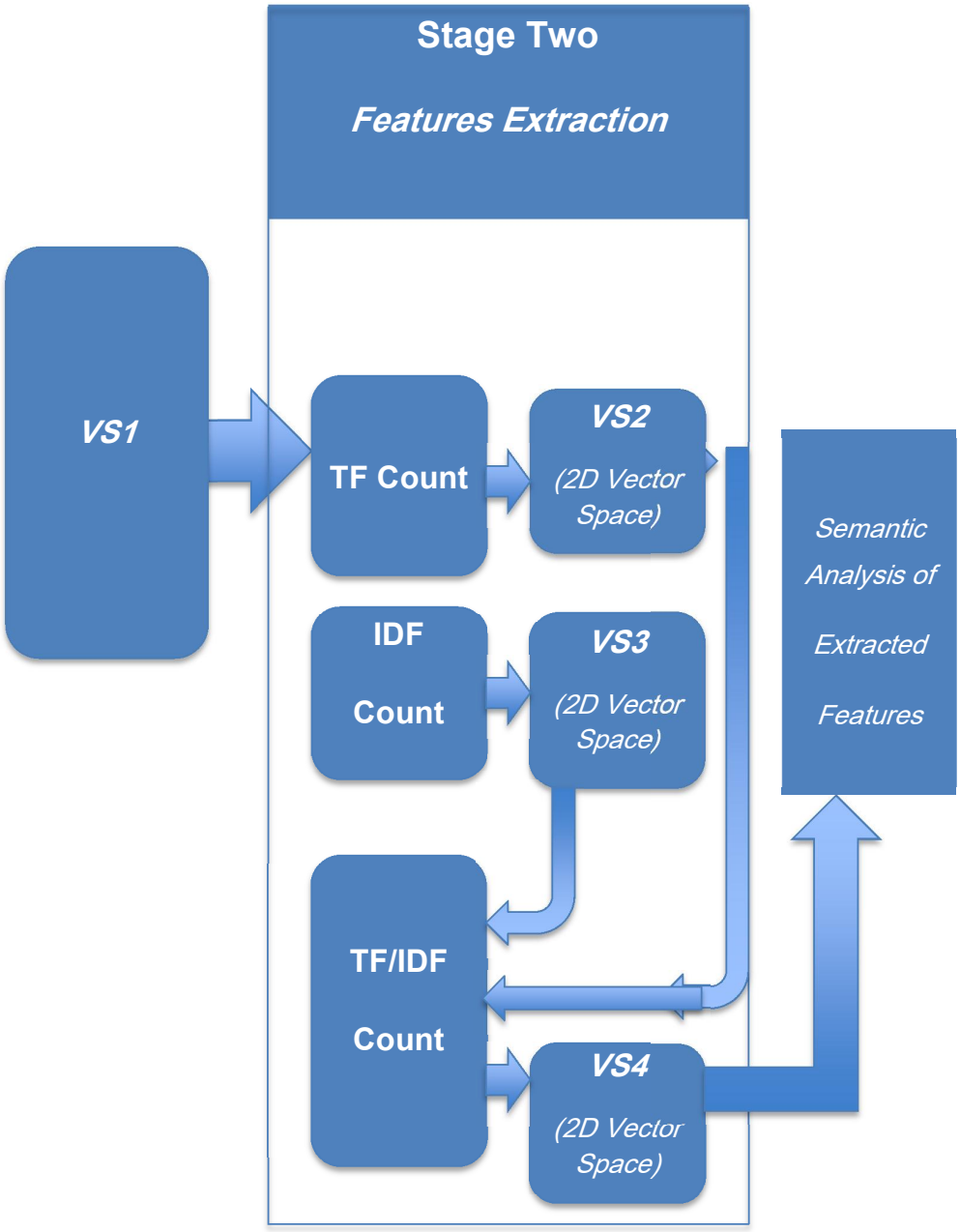

Figure 4.4 Features Extraction Stage

Table 4.3 Phase 1 Classification Performance of the RF Classifier for Thresholds > 1, 2, 3 and 4

Threshold >	Number of reduced features	Cross Validation (Folds)	Accuracy %
1	3210	5	16.7702
		10	83.2298
		20	85.7143
		30	***85.7143***
		40	85.7143
		50	84.472
2	***709***	5	81.3665
		10	78.882
		20	85.0932
		30	***88.8199***
		40	84.472
		50	82.6087
3	283	***5***	***88.1988***
		10	85.0932
		20	83.8509
		30	86.3354
		40	85.7143
		50	85.0932
4	158	5	78.2609
		10	80.1242
		20	79.5031
		30	***81.9876***
		40	80.7453
		50	78.882

Table 4.4 Phase 1 Classification Performance of the RF Classifier for Thresholds > 5, 6, 7 and 8

Threshold >	Number of reduced features	Cross Validation (Folds)	Accuracy %
5	88	5	74.5342
		10	83.2298
		20	81.3665
		30	78.882
		40	80.7453
		50	***83.8509***
6	60	5	77.6398
		10	76.3975
		20	80.1242
		30	79.5031
		40	81.3665
		50	***82.6087***
7	42	5	77.6398
		10	77.0186
		20	79.5031
		30	***80.1242***
		40	77.6398
		50	78.882
8	30	5	68.9441
		10	68.323
		20	69.5652
		30-40	***70.8075***
		50	69.5652

Table 4.5 Phase 1 Classification Performance of the RF Classifier for Thresholds > 9, 10, 11, 12, 13, 14 and 15

Threshold >	Number of reduced features	Cross Validation (Folds)	Accuracy %
9	19	5	69.5652
		10	68.323
		20	70.1863
		30	***72.6708***
		40	72.0497
		50	71.4286
10	14	***5-10-20-30-40-50***	***72.0497***
11	13	***5-10-20-30-40-50***	***72.0497***
12	12	***5-10-20-30-40-50***	***72.0497***
13	9	***5-10-20-30-40-50***	***72.0497***
14	7	***5-10-20-30-40-50***	***72.0497***
15	5	***5-10-20-30-40-50***	***70.8075***

Table 4.6 Phase 2-Experiment 1 Classification Performance of the RF Classifier for Thresholds > 1, 2, 3 and 4

Threshold >	Number of reduced features	Cross Validation (Folds)	Accuracy %
1	4074	5	89.8897
		10	96.3235
		20	97.6103
		30	95.9559
		40	***98.1618***
		50	97.2426
2	***1057***	5	90.9926
		10	96.3235
		20-30	97.2426
		40	***98.3456***
		50	98.1618
3	456	5	90.0735
		10	94.3015
		20	97.4265
		30	97.2426
		40	***97.6103***
		50	97.4265
4	261	5	89.7059
		10	95.7721
		20	97.2426
		30	96.5074
		40	***97.9779***
		50	97.6103

Table 4.7 Phase 2-Experiment 1 Classification Performance of the RF Classifier for Thresholds > 5, 6, 7 and 8

Threshold >	Number of reduced features	Cross Validation (Folds)	Accuracy %
5	159	5	89.5221
		10	93.75
		20	97.4265
		30	96.875
		40	***97.9779***
		50	97.7941
6	113	5	90.0735
		10	93.75
		20	95.0368
		30	95.4044
		40	***96.3235***
		50	95.7721
7	88	5	89.3382
		10	91.1765
		20	95.0368
		30	94.6691
		40	***96.1397***
		50	95.4044
8	64	5	88.6029
		10	92.2794
		20	93.9338
		30	94.3015
		40	95.4044
		50	***95.5882***

Table 4.8 Phase 2-Experiment 1 Classification Performance of the RF Classifier for Thresholds > 9, 10, 11 and 12

Threshold >	Number of reduced features	Cross Validation (Folds)	Accuracy %
9	47	5	88.6029
		10	91.3603
		20	94.1176
		30	94.6691
		40	***95.9559***
		50	95.5882
10	38	5	83.8235
		10	87.6838
		20	88.7868
		30	89.7059
		40	90.4412
		50	***90.8088***
11	32	5	80.6985
		10	84.5588
		20	86.3971
		30	85.6618
		40	***87.8676***
		50	87.6838
12	27	5	70.0368
		10	72.4265
		20	73.7132
		30	72.9779
		40	***74.2647***
		50	73.8971

Table 4.9 Phase 2-Experiment 1 Classification Performance of the RF Classifier for Thresholds > 13, 14 and 15

Threshold >	Number of reduced features	Cross Validation (Folds)	Accuracy %
13	23	5	61.0294
		10	62.8676
		20	63.6029
		30	63.2353
		40	***63.7868***
		50	62.3162
14	18	5	55.6985
		10	58.8235
		20-30	59.0074
		40	***59.5588***
		50	59.1912
15	15	5	53.3088
		10	55.3309
		20	55.1471
		30	55.5147
		40	***55.6985***
		50	55.5147

Table 4.10 Phase 2-Experiment 2 Classification Performance of the RF Classifier for Thresholds > 1, 2, 3 and 4

Threshold >	Number of reduced features	Cross Validation (Folds)	Accuracy %
1	***5987***	5	91.5441
		10	95.4044
		20	97.4265
		30	95.7721
		40-50	***98.8971***
2	667	5	85.2941
		10	91.9118
		20	***93.5662***
		30	91.1765
		40	92.8309
		50	91.7279
3	242	5	72.0588
		10	72.4265
		20	73.1618
		30	***75.1838***
		40	74.6324
		50	74.6324
4	107	5	60.8456
		10	61.7647
		20	62.5
		30	62.8676
		40	***63.4191***
		50	61.7647

Table 4.11 Phase 2-Experiment 2 Classification Performance of the RF Classifier for Thresholds > 5, 6, 7 and 8

Threshold >	Number of reduced features	Cross Validation (Folds)	Accuracy %
5	72	5	56.4338
		10	56.8015
		20-30-40	57.3529
		50	***57.7206***
6	52	5	51.2868
		10	52.0221
		20	51.6544
		30	***52.3897***
		40-50	52.0221
7	34	***5***	***50.9191***
		10	50.7353
		20	50
		30	50.5515
		40	50.3676
		50	50.7353
8	26	***5***	***49.4485***
		10	49.2647
		20	48.5294
		30	49.0809
		40	48.8971
		50	49.2647

Table 4.12 Phase 2-Experiment 2 Classification Performance of the RF Classifier for Thresholds > 9, 10 and 11

Threshold >	Number of reduced features	Cross Validation (Folds)	Accuracy %
9	22	***5***	***49.0809***
		10	48.8971
		20	48.1618
		30	48.7132
		40	48.5294
		50	48.8971
10	21	***5***	***48.1618***
		10	47.9779
		20	47.7941
		30	47.6103
		40-50	***48.1618***
11	19	***5***	***46.6912***
		10	46.5074
		20	46.3235
		30	46.1397
		40-50	***46.6912***

4.4 Semantic Analysis of Extracted Features (Stage three)

This section discusses the semantic analysis of extracted features, executed on ***VS4*** produced by the end of stage two. This stage aims at discovering the relationships between these features and classifying them in addition to the news articles into appropriate semantic classes (Figure 4.5).

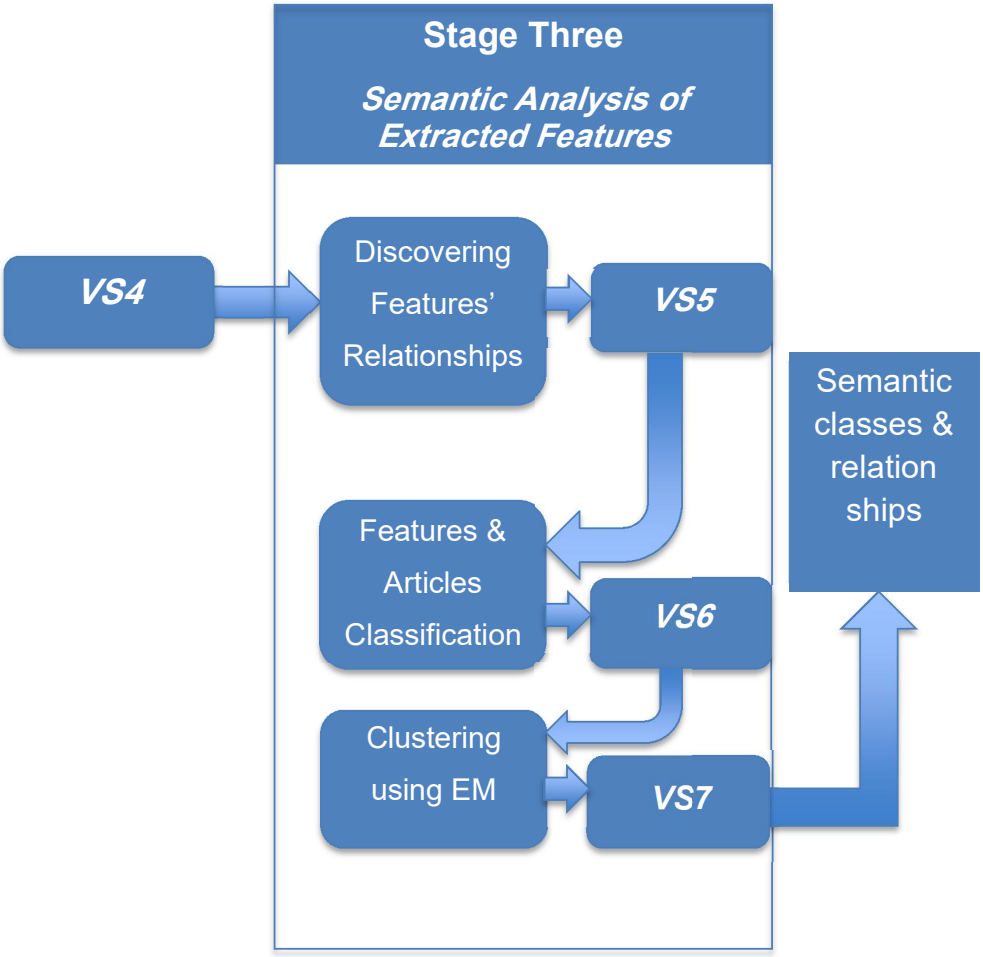

Figure 4.5 SMRF-TM Stage Three

4.4.1 Application of Random Forest

Stage three of SMRF-TM aims at discovering the relationships between the extracted features and classifying these features into one of the five classes: critical down, down, neutral, up and critical up, defined in section 4.5. This is one of the novel contributions of the SMRF-TM approach since most of the previous text mining techniques used three classes only (down, neutral and up) to classify the news articles related to the stock market domain as explained in the literature review chapter.

In stage three of SMRF-TM, RF is used to reveal the hidden information and relations between the extracted features in (***VS4***) by generating a random forest consisting of 10 random trees, each constructed while considering a number (*n*) of random features. To classify a new feature/news article from the input features set/news corpus, the input set of extracted features/news articles is placed in each of the trees in the forest. Each tree then gives a classification and the forest chooses the classification having the majority votes among the 10 trees in order to classify the new feature/article into one of the above five classes. The final output of this task is ***VS6***, which serves as the input for the next clustering task using expectation maximisation, described below.

The random forest application in the SMRF-TM approach generates an ensemble of 10 random, individual and un-pruned trees. Each individual tree is constructed using the following algorithm:

Random Forest Pseudo code

Inputs: t (the number of random trees in the forest (iterations = 10))

S (the training set)

n (number of random features used in constructing each of the 10 trees)

Outputs: T_t ; t =1,…., 10

1) t = 1
2) Do
3) s_t is a subsample articles from S with replacement
4) Construct classifier tree T_t using a decision tree inducer on s_t
5) t ++
6) while (t < 10)

The input parameter (n) represents the number of features, which is used to determine the decision at a node of the tree and it should be much less than the total number of features in the training set (S). The constructed ensemble decision trees (10 trees) are not pruned and the best split at each node is chosen from among the (n) random features not all the features. The classification of any unlabelled feature/news article is performed using the majority votes.

Figures 4.6 and 4.7 show samples of how each tree in the RF discovers the relationships between the unigrams and bigrams features respectively in the implementation of SMRF-TM.

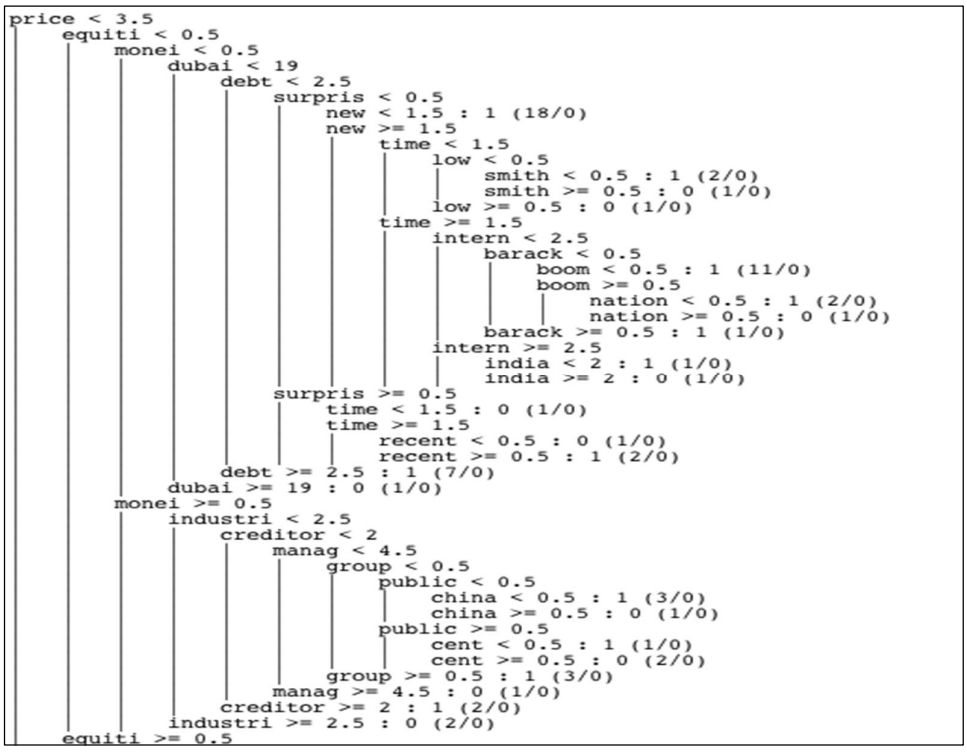

Figure 4.6 Sample of how RF discovers the relationships between unigrams features in the SMRF-TM approach

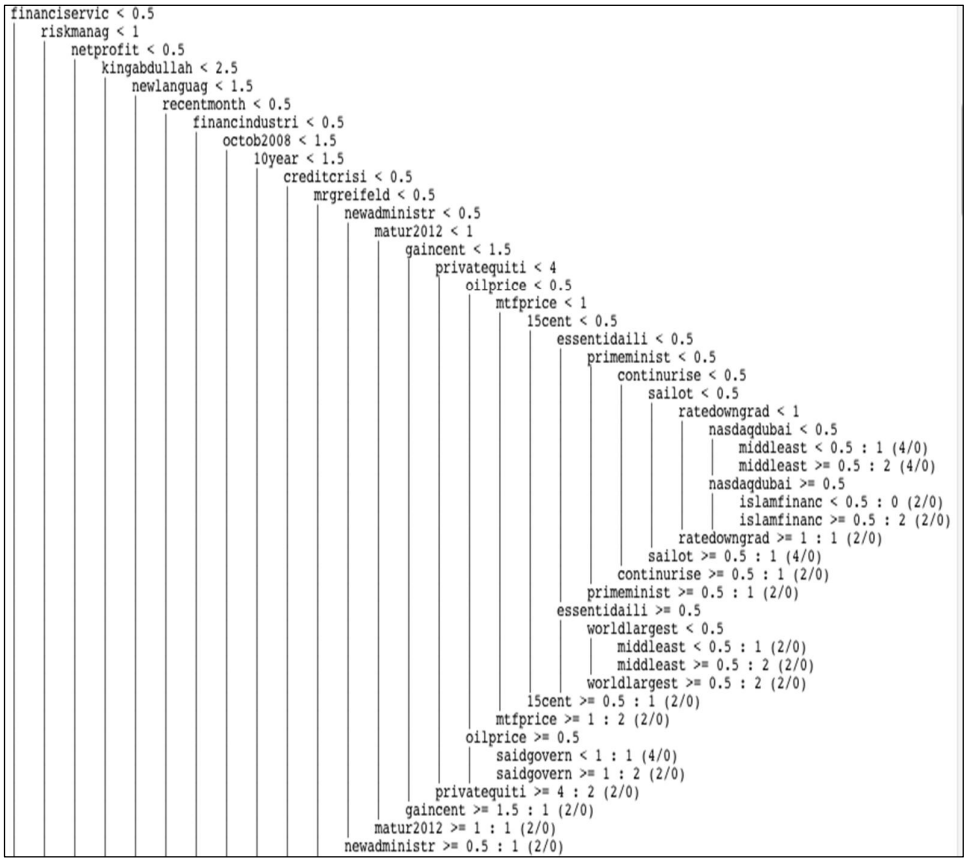

Figure 4.7 Sample of how RF discovers the relationships between bigrams features in the SMRF-TM approach

WEKA software is used to apply the RF classifier to classify the extracted features/news articles based on ***VS4***. By setting a threshold > 2, the classification based on TF/IDF, has reduced 9,198 features to 709 features in phase one, from 11,036 features to 1057 features in phase two experiment one and in phase two experiment two the features were reduced from 82,814 features to 5,987 features by setting a threshold > 1. The threshold is changed from >2 to >1 in experiment two of phase two because analysing bigram tokens is more efficient with bigger datasets and so the best classification accuracy achieved by the RF is 98.89% when threshold is set to >1 as shown above in Tables 4.10, 4.11 and 4.12.

These Random forest results are compared with other different classifiers such as ADTree, J48, J48graft, Decision Stump, Random Tree, Bayes Net, Bagging, Rotation Forest and Decision Table. The results of these comparisons are discussed later in section 4.5. Each type of these classifiers has different characteristics, which can significantly affect the performance of the SMRF-TM approach. For example, Bayes classifiers are composed of directed acyclic graphs with only one parent and several children and they assume that child nodes are independent in order to simplify learning. Even though independence is considered as a poor and unrealistic assumption, Bayes classifiers may still compete sometimes with more sophisticated classifiers because of its short computational time for training (Kotsiantis *et al.* 2007, Rish 2001).

On the other hand, trees classifiers depend on the features values to classify articles by sorting them according to these values, where each node in a tree represents a feature in an instance to be classified, and each branch represents a value, which the node can assume. Starting at the root node articles are classified and sorted according to their features values.

In order to validate the results and obtain performance accuracy, cross validation with different folds (5, 10, 20, 30, 40, 50) are used to check which classifier has the best learning capabilities to achieve the best classification performance. In cross validation, the training set is divided into mutually exclusive and equal sized subsets and for each subset the classifier is trained on the union of all the other subsets. The average of the error rate of each subset is therefore an estimate of the error rate of the classifier (Kotsiantis 2007). For example, the 10-folds cross validation uses 9/10 of the data for training the algorithm and 1/10 of the data for testing, then repeats this process 10 times after shuffling the data each time.

In phase one, only two classes are implemented (up and down) in order to evaluate the proof of concept. In phase two experiment one, the classes were extended to include the five classes: critical down, down, neutral, up and critical up, this has significantly enhanced the classification performance of RF from 88.82% to 98.34%. In phase two experiment two, the classes are also extended, but the extracted features, which are used to support the classification process are bigram features, not unigram features as in phase one and phase two experiment one.

4.4.2 Application of Expectation Maximisation

"The Expectation Maximisation (EM) clustering algorithm is considered as an appropriate optimisation algorithm for constructing proper statistical models of the data, which admits varying degrees of data membership in multiple clusters. EM is an effective, popular technique for estimating mixture model parameters. The EM algorithm iteratively refines initial mixture model parameter estimates to better fit the data and terminates at a locally optimal solution. EM has been shown to be superior to other alternatives for statistical modelling purposes" (Bradley *et al.* 2000).

In stage three of SMRF-TM the EM is applied on ***VS6*** using also WEKA to cluster the classified features and the news articles according to their semantic meanings in one of the three clusters: economic, social or political. The output of this task is the ***VS7***, which is the final output of SMRF-TM. The ***VS7*** can then be used to enhance the accuracy of predicting Dubai's SM movements. This is not only able to classify the features/articles according to the predicted influence they have on Dubai's SM movements but is also able to examine the reason behind such movements.

Tables 4.13 and 4.14 show samples of the clustered unigrams and bigrams features according to the three above clusters, which are then used to cluster the news articles accordingly.

Table 4.13 Sample of clustered unigram features

Economic	Social	Political
fund	korea	govern
invest	china	vote
return	saudi	public
low	dubai	plan
higher	middl	polici
rate	east	parliament
increas	cultur	candid
budget	chines	democraci
bond	europ	law
Stock	india	council
market	nation	polit
investor	peopl	elect
financi	countri	regim
bank	london	conserv
growth	uk	diplomat
sharehold	russia	lawyer
earn	gulf	mayor
spend	european	governor

Table 4.14 Sample of clustered bigram features

Economic	Social	Political
britaineconom	nigerdelta	pressfreedom
ukeconom	visitvatican	policitighten
globalbond	peoplconcern	parliamentresult
inflationaririsk	peoplworri	hungparliament
debtsustain	peoplstop	democraciactivist
bigrisk	concernpeopl	uaedemocraci
riskuk	nationinsur	conservwin
priceinflat	worldclass	governwant
helpeconomi	nationinfrastructur	policiframework
riskeconomi	helpfamili	conservgovern
pricetarget	helppeopl	weakgovern
advanccent	britainstrong	policicontinu
highpaid	generexpect	legalservic
sectorinvest	riseunemploy	lawpublic
debtshare	uklike	founderlaw
incomtax	expectuk	counticouncil
promotgrowth	likelihoodsignific	directorgovern
taxcredit	londonschool	lawpartnership
valufair	chinaproperti	youvote
corportax	clearhous	theyvote
investvital	allianzkorea	newgovern

4.5 Discussion

The five classes used in this research represent the predicted influence of the features/articles on Dubai's SM movements. In phase one, the experiment was limited to two categories. The down class, which means that they have a small negative influence on Dubai's SM causing a decrease less than 4% in the value of Dubai's SM general index (DFM) and the up class, which means that they have a small positive influence on Dubai's SM resulting in an increase less than 4% in the value of Dubai's SM general index (DFM). Three more classes are added in phase two: the critical down/up classes, which means that they have a big negative/positive influence on Dubai's SM causing a decrease/increase more than 4% in the value of Dubai's SM general index (DFM) and the neutral class, which means that they do not have any influence on Dubai's SM general index (DFM).

4.5.1 Results of Phase One

The classification results in Tables 4.15, 4.16 and 4.17, below, show that the ADTree, RF, Bagging, J48graft, Bayes Net, and Decision Table classifiers outperformed the J48, Decision Stump and Random Tree classifiers. However, the Rotation Forest classifier ran out of memory. The ADTree and the RF classifiers, which yielded the best classification performance among the 10 tested classifiers, achieved the highest accuracy by deploying the 30-folds cross validation.

The results are expressed in terms of precision, recall and accuracy defined as follow:

$$Precision\ (PR) = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Positive\ (FP)} \quad (4.1)$$

$$Recall\ (RC) = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Negative\ (FN)} \quad (4.2)$$

$$Accuracy = \frac{Total\ number\ of\ correctly\ classified\ instances}{Total\ number\ of\ instances} * 100 \quad (4.3)$$

$$Fmeasure = \frac{2 * PR * RC}{PR + RC} \quad (4.4)$$

Where;

True Positive (TP): measures the proportion of positives, which are correctly identified as positives.

True Negative (TN): measures the proportion of negatives, which are correctly identified as negatives.

False Positive (FP): measures the proportion of negatives, which are incorrectly identified as positives.

False Negative (FN): measures the proportion of positives, which are incorrectly identified as negatives.

Out of 161 articles, 114 belong to the down class and 47 to the up class. By analysing the results of the best three classifiers (ADTree, Random Forest and Bagging), we found that the ADTree classifier has achieved classification accuracy 91.92% with a precision 0.917 and recall 0.974 for the down class but with a higher precision 0.925 and lower recall 0.787 for the up class. The reasons behind these results are: the ADTree classifier has classified 121 articles as down instead of 114 and 40 articles as up instead of 47. Out of the 121 articles, which are classified as down; 111 articles true positive, 10 articles false positive and three articles false negative. Out of the 40 articles, which are classified as up; 37 articles true positive, three articles false positive and 10 articles false negative. The total number of incorrectly classified articles is 13 out of 161 articles and most of them (10 out of 13) belong to the up class.

Regarding the RF, the classification accuracy achieved is 88.82% with a precision 0.887 and recall 0.965 for the down class but with a higher precision 0.892 and lower recall 0.702 for the up class. This is because the RF has correctly classified 110 articles out of 114 for the down class, which means that, the down class has 110 articles true positive, 14 articles false positive and four articles false negative. Concerning the up class, RF has correctly classified 33 articles out of 47 indicating that the up class has 33 articles true positive, four articles false positive and 14 articles false negative.

Table 4.15 Phase 1 the RF, ADTRee and J48 Classification Performance for Threshold > 2

Classifier	Cross Validation (Folds)	Accuracy %	Precision (Down class)	Recall (Down class)	Precision (Up class)	Recall (Up class)
RF	5	81.3665	0.818	0.947	0.793	0.489
	10	78.882	0.803	0.93	0.724	0.447
	20	85.0932	0.852	0.956	0.848	0.596
	30	***88.8199***	***0.887***	***0.965***	***0.892***	***0.702***
	40	84.472	0.85	0.947	0.824	0.596
	50	82.6087	0.831	0.947	0.806	0.532
ADTree	5	85.7143	0.87	0.939	0.816	0.66
	10	86.3354	0883	0.93	0.805	0.702
	20	88.1988	0.886	0.956	0.868	0.702
	30	***91.9255***	***0.917***	***0.974***	***0.925***	***0.787***
	40	90.0621	0.902	0.965	0.897	0.745
	50	89.441	0.888	0.974	0.917	0.702
J48	***5***	***85.0932***	***0.875***	***0.921***	***0.78***	***0.681***
	10	83.2298	0.866	0.904	0.738	0.66
	20	81.9876	0.863	0.886	0.705	0.66
	30	83.2298	0.866	0.904	0.738	0.66
	40	82.6087	0.858	0.904	0.732	0.638
	50	83.8509	0.867	0.912	0.756	0.66

Table 4.16 Phase 1 the J48graft, Decision Stump, Random Tree and Bayes Net Classification Performance for Threshold > 2

Classifier	Cross Validation (Folds)	Accuracy %	Precision (Down class)	Recall (Down class)	Precision (Up class)	Recall (Up class)
J48graft	5	86.3354	0.877	0.939	0.821	0.681
	10	***87.5776***	***0.885***	***0.947***	***0.846***	***0.702***
	20	86.9565	0.884	0.939	0.825	0.702
	30	86.3354	0.877	0.939	0.821	0.681
	40	86.3354	0.877	0.939	0.821	0.681
	50	85.7143	0.87	0.939	0.816	0.66
Decision Stump	5	83.8509	0.844	0.947	0.818	0.574
	10	83.8509	0.844	0.947	0.818	0.574
	20	83.8509	0.844	0.947	0.818	0.574
	30	***83.8509***	***0.844***	***0.947***	***0.818***	***0.574***
	40	83.8509	0.844	0.947	0.818	0.574
	50	83.8509	0.844	0.947	0.818	0.574
Random Tree	5	76.3975	0.797	0.895	0.636	0.447
	10	77.0186	0.813	0.877	0.632	0.511
	20	78.882	0.828	0.886	0.667	0.553
	30	79.5031	0.829	0.895	0.684	0.553
	40	79.5031	0.829	0.895	0.684	0.553
	50	***80.1242***	***0.836***	***0.895***	***0.692***	***0.574***
Bayes Net	***5***	***87.5776***	***0.879***	***0.956***	***0.865***	***0.681***
	10	83.2298	0.854	0.921	0.763	0.617
	20	83.2298	0.848	0.93	0.778	0.596
	30	80.7453	0.832	0.912	0.722	0.553
	40	81.9876	0.846	0.912	0.737	0.596
	50	81.9876	0.846	0.912	0.737	0.596

Table 4.17 Phase 1 the Bagging, Rotation Forest and Decision Table Classification Performance for Threshold > 2

Classifier	Cross Validation (Folds)	Accuracy %	Precision (Down class)	Recall (Down class)	Precision (Up class)	Recall (Up class)
Bagging	5	85.0932	0.869	0.93	0.795	0.66
	10	86.9565	0.891	0.93	0.81	0.723
	20	86.9565	0.891	0.93	0.81	0.723
	30	86.3354	0.883	0.93	0.805	0.702
	40	***88.1988***	***0.899***	***0.939***	***0.833***	***0.745***
	50	87.5776	0.905	0.921	0.8	0.766
Rotation Forest	5	85.7143	0.858	0.956	0.853	0.617
	10	89.441	0.894	0.965	0.895	0.723
	20	86.9565	0.891	0.93	0.81	0.723
	30	89.441	0.915	0.939	0.841	0.787
	40	90.0621	0.908	0.956	0.878	0.766
	50	***90.6832***	***0.916***	***0.956***	***0.881***	***0.787***
Decision Table	***5***	***87.5776***	***0.898***	***0.93***	***0.814***	***0.745***
	10	83.8509	0.879	0.895	0.733	0.702
	20	85.7143	0.889	0.912	0.773	0.723
	30	84.472	0.88	0.904	0.75	0.702
	40	85.7143	0.889	0.912	0.773	0.723
	50	86.9565	0.897	0.921	0.795	0.745

The RF misclassified five more articles than the ADTree with a total number of 18 incorrectly classified articles, which indicates that the ADTree is more efficient than the RF when the dataset is small. But still most of the incorrectly classified articles (14 out of 18) belong to the up class.

The Bagging classifier achieved the best classification performance while deploying 40-folds cross validation, which is 88.19% with a precision 0.899 and recall 0.939 for the down class but with a lower precision 0.833 and lower recall 0.745 for the up class. The Bagging classifier yielded these results because it has correctly classified 107 articles out of 114 for the down class and 35 articles out of 47 for the up class, with total number of 19 incorrectly classified articles, 12 of them belong to the up class.

This shows that the tree classifiers (ADTree and RF) perform better than the meta classifiers such as Bagging, when applied on large dataset (down class) as the ADTree misclassified three articles out of 114 with precision 0.917 and recall 0.974; the RF had four incorrectly classified articles out of 114 with a precision 0.887 and recall 0.965 while Bagging had seven incorrectly classified articles with a precision 0.899 and recall 0.939 out of 114 articles, which belong to the down class. In addition, it shows that the RF classifier had the worst performance on the small dataset (up class) as it had 14 incorrectly classified articles, Bagging had 12 incorrectly classified articles while the ADTree had only 10 incorrectly classified articles out of 47 articles, which belong to the up class.

As mentioned above the majority of the incorrectly classified articles among the three classifiers belong to the up class as the dataset used so far for this class consists only of 47 articles, which was obviously not enough to extract the desired discriminative features for this class.

The application of the EM clustering technique in this phase of the implementation resulted in clustering the majority of the extracted features to the economic cluster; 93% of the news articles were clustered to the economic cluster, 5% of the news articles to the social cluster and 2% of the news articles to the political cluster. Such results are considered reasonable because the source of our dataset is Financial Times, which publishes only financial news.

The classification results for threshold > 3 shown below in Tables 4.18, 4.19 and 4.20, clarify that the classification performances of the RF, ADTree and the Bayes Net classifiers are decreased as the number of the extracted features is decreased from 709 to 283 features. The classification performances of the J48, J48graft, Bagging and Decision Table classifiers are increased when the number of features decreased while for the rest of the classifiers tested the classification performance almost remained unchanged. This indicates that RF classifier performs better than the other classifiers when applied on large datasets. The effect of enlarging the dataset is checked in the two experiments of phase two of the implementation. The results also show that the RF, ADTree, J48graft, Bagging, Rotation Forest and Decision Table classifiers outperformed the J48, Decision Stump, Random Tree and Bayes Net classifiers. For most of the tested classifiers the cross validation with 5-folds yielded the best results except for the ADTree and Random Tree classifiers, which indicates that most of the above classifiers perform better when the training dataset increases.

Comparing the performance of the best six classifiers, with accuracy ranging between 88.19% and 88.82%, the three classifiers: RF, J48graft and ADTree classifiers have 19 incorrectly classified articles resulting in exactly the same classification performance 88.19% but with different precisions and recalls for the down and the up classes.

The RF has correctly classified 113 articles out of 114 articles for the down class and 29 articles out of 47 articles for the up class, which means that the down class has 113 articles true positive, 18 articles false positive and one article false negative, while the up class has 29 articles true positive, one article false positive and 18 articles false negative. So, the down class has precision 0.863 and recall 0.991 but the up class has higher precision 0.967 and lower recall 0.617. Most of the incorrectly classified articles 18 out of 19 articles belong to the up class.

The J48graft has correctly classified 107 articles out of 114 articles for the down class and 35 articles out of 47 articles for the up class. Consequently, the down class has 107 articles true positive, 12 articles false positive and seven articles false negative and the up class has 35 articles true positive, seven articles false positive and 12 articles false negative. Hence, the down class has precision 0.899 and recall 0.939 but the up class has lower precision 0.833 and lower recall 0.745. The number of the incorrectly classified articles belonging to the down class increased from one article to seven articles and decreased for the up class from 18 articles to 12 articles.

The ADTree correctly classified 108 articles out of 114 articles for the down class resulting in 0.893 precision and 0.947 recall. Because the down class has 108 articles true positive, 13 articles false positive and six articles false negative. As for the up class, the ADTree correctly classified 34 articles out of 47 articles with 34 articles true positive, six articles false positive and 13 articles false negative, which yielded lower precision 0.85 and lower recall 0.723. Again, the number of the incorrectly classified articles belonging to the down class increased from one to six and decreased for the up class from 18 to 13. These results indicate that the RF performs better and is more precise than the J48graft and the ADTree when applied on large datasets (down class) and the contrary for small datasets (up class).

Table 4.18 Phase 1 the RF, ADTree and J48 Classification Performance for Threshold > 3

Classifier	Cross Validation (Folds)	Accuracy %	Precision (Down class)	Recall (Down class)	Precision (Up class)	Recall (Up class)
RF	***5***	***88.1988***	***0.863***	***0.991***	***0.967***	***0.617***
	10	85.0932	0.846	0.965	0.871	0.574
	20	83.8509	0.849	0.939	0.8	0.596
	30	86.3354	0.854	0.974	0.903	0.596
	40	85.7143	0.847	0.974	0.9	0.574
	50	85.0932	0.836	0.982	0.926	0.532
ADTree	5	85.7143	0.903	0.895	0.75	0.766
	10	83.8509	0.873	0.904	0.744	0.681
	20	83.8509	0.867	0.912	0.756	0.66
	30	***88.1988***	***0.893***	***0.947***	***0.85***	***0.723***
	40	85.0932	0.875	0.921	0.78	0.681
	50	85.0932	0.881	0.912	0.767	0.702
J48	***5***	***86.9565***	***0.897***	***0.921***	***0.795***	***0.745***
	10	86.3354	0.89	0.921	0.791	0.723
	20	83.8509	0.879	0.895	0.733	0.702
	30	85.0932	0.902	0.886	0.735	0.766
	40	81.9876	0.876	0.868	0.688	0.702
	50	85.0932	0.888	0.851	0.756	0.723

Table 4.19 Phase 1 the J48graft, Decision Stump and Random Tree Classification Performance for Threshold > 3

Classifier	Cross Validation (Folds)	Accuracy %	Precision (Down class)	Recall (Down class)	Precision (Up class)	Recall (Up class)
J48graft	***5***	***88.1988***	***0.899***	***0.939***	***0.833***	***0.745***
	10	86.9565	0.891	0.93	0.81	0.723
	20	86.3354	0.883	0.93	0.805	0.702
	30	85.7143	0.896	0.904	0.761	0.745
	40	83.8509	0.879	0.895	0.733	0.702
	50	85.7143	0.889	0.912	0.773	0.723
Decision Stump	***5***	***83.8509***	***0.844***	***0.947***	***0.818***	***0.574***
	10	83.8509	0.844	0.947	0.818	0.574
	20	83.8509	0.844	0.947	0.818	0.574
	30	83.8509	0.844	0.947	0.818	0.574
	40	83.8509	0.844	0.947	0.818	0.574
	50	83.8509	0.844	0.947	0.818	0.574
Random Tree	5	79.5031	0.809	0.93	0.733	0.468
	10	74.5342	0.797	0.86	0.579	0.468
	20	***81.9876***	***0.846***	***0.912***	***0.737***	***0.596***
	30	80.1242	0.811	0.939	0.759	0.468
	40	77.6398	0.815	0.886	0.649	0.511
	50	75.7764	0.815	0.851	0.595	0.532

Table 4.20 Phase 1 the Bayes Net, Bagging, Rotation Forest and Decision Table Classification Performance for Threshold > 3

Classifier	Cross Validation (Folds)	Accuracy %	Precision (Down class)	Recall (Down class)	Precision (Up class)	Recall (Up class)
Bayes Net	*5*	*86.3354*	*0.883*	*0.93*	*0.805*	*0.702*
	10	85.7143	0.876	0.93	0.8	0.681
	20	81.9876	0.851	0.904	0.725	0.617
	30	82.6087	0.864	0.895	0.721	0.66
	40	84.472	0.862	0.93	0.789	0.638
	50	81.3665	0.844	0.904	0.718	0.596
Bagging	*5*	*88.8199*	*0.907*	*0.939*	*0.837*	*0.766*
	10	85.7143	0.876	0.93	0.8	0.681
	20	85.7143	0.889	0.912	0.773	0.723
	30	87.5776	0.912	0.912	0.787	0.787
	40	87.5776	0.912	0.912	0.787	0.787
	50	86.9565	0.904	0.912	0.783	0.766
Rotation Forest	*5*	*88.8199*	*0.907*	*0.939*	*0.837*	*0.766*
	10	88.8199	0.9	0.947	0.854	0.745
	20	88.8199	0.899	0.939	0.833	0.745
	30	87.5776	0.898	0.93	0.814	0.745
	40	86.3354	0.897	0.912	0.778	0.745
	50	86.9565	0.891	0.93	0.81	0.723
Decision Table	*5*	*88.8199*	*0.914*	*0.93*	*0.822*	*0.787*
	10-20-30-40-50	87.5776	0.905	0.921	0.8	0.766

The Bagging, Rotation Forest and Decision Table classifiers have only 18 incorrectly classified articles with the same classification performance 88.82%. The Bagging and the Rotation Forest correctly classified 107 articles out of 114 for the down class and 36 articles out of 47 for the up class. Hence, the precision for the down class is 0.907 and the recall is 0.939 as it has 107 articles true positive, 11 articles false positive and seven articles false negative. But the up class has lower precision 0.837 and lower recall 0.766 because it has 36 articles true positive, seven articles false positive and 11 articles false negative. Most of the incorrectly classified articles 11 out of 18 belong to the up class.

The Decision Table correctly classified 106 articles out of 114 for the down class and 37 articles out of 47 for the up class. Accordingly, the down class has 106 articles true positive, 10 articles false positive and eight articles false negative resulting in 0.914 precision and 0.93 recall. The up class has 37 articles true positive, eight articles false positive and 10 articles false negative resulting in lower precision 0.822 and lower recall 0.787.

The number of incorrectly classified articles belongings to the down class increased from seven to eight and decreased for the up class from 11 to 10. This indicates that the Bagging and the Rotation Forest perform better than the Decision Table when applied on large datasets and the contrary for small datasets.

RF achieved the best performance among the six classifiers when applied on the bigger dataset (down class). However, the majority of the incorrectly classified articles among all of the six classifiers belong to the up class. As already noted, the small number of articles (47) belonging to the up class in our dataset has negatively affected the performance in phase one of the implementation, hence, this problem is addressed in phase two (experiment one and experiment two) through the expansion of the dataset especially for the up class.

The incorrectly classified articles yielded by the tested classifiers are incorrectly classified based on features that should belong to the neutral class such as time, world, year, decade, like, term, public, alkhaleej, Dubai, China, UK, automobile, Nigerian, ship, food, school, and other similar features. Consequently, to improve the performance of the classifiers in phase two of the implementation of SMRF-TM the two classes (down and up) used in phase one are extended into five classes (critical down, down, neutral, up and critical up) as shown in the following sections.

Examples of the extracted classified features used to support in the classification of the news articles in phase one is available in Appendix A also showing the features, which were classified to more than one class.

4.5.2 Results of Experiment One of Phase Two

Most of the former researchers in this field, as discussed in the literature review, deployed only three classes: down, neutral and up, to classify news articles. The first experiment of phase two applied the extended classes (i.e. critical down, down, neutral, up and critical up) to classify the extracted features from the dataset, which are then used to classify the 544 articles accordingly. This helps us examine the effect of expanding the dataset and the classification classes on the performance of our SMRF-TM approach.

By comparing the classification results of experiment one in phase two (Tables 4.21, 4.22, 4.23, 4.24 and 4.25) with the classification results of phase one (Tables 4.15, 4.16 and 4.17) we find that the classification performance of both, the RF and the Random Tree, is notably enhanced from 88.82% to 98.34% and from 80.12% to 98.16% respectively, while for the rest of the tested classifiers their performances were notably decreased. Furthermore, RF and Random Tree classifiers outperformed all the other tested classifiers. They have achieved the highest accuracy by deploying the 40-folds cross validation. This means that, given the

expanded dataset size we could use 60% of the dataset (326 articles) for training rather than 70% (113 articles) for the 30-folds cross validation used in phase one as the number of articles used for training is increased.

Out of 544 articles, 134 articles are attributed to the neutral class, 184 articles to the down class, 138 articles to the up class, 54 articles to the critical down class and 34 articles to the critical up class. By analysing the results of the best two classifiers it was found that the RF classifier has correctly classified 535 out of 544 of the dataset corpus resulting in 98.34% classification accuracy. The total number of the incorrectly classified articles using the RF classifier is 9 articles out of 544, which are distributed as follows: two articles out of 134 for the neutral class resulting in 132 articles true positive, two articles false positive and two articles false negative. The down class has two incorrectly classified articles out of 184 yielding 182 articles true positive, two articles false positive and two articles false negative. The up class has four incorrectly classified articles out of 138 for producing 134 articles true positive, five articles false positive and four articles false negative. The critical down class has 54 articles correctly classified so it has 54 articles true positive, zero articles false positive and zero articles false negative. Finally, the critical up class has one incorrectly classified article out of 34, which means that it has 33 articles true positive, zero article false positive and one article false negative.

The distribution of the correctly classified articles, precision and recall for the five classes are as follows: 132 articles out of 134 for the neutral class with 0.985 precision and recall, 182 articles out of 184 for the down class with 0.989 precision and recall, 134 articles out of 138 for the up class with 0.964 precision and 0.971 recall, 54 articles out of 54 for the critical down class with 1.000 precision and recall and for the critical up class 33 articles out of 34 are correctly classified with 1.000 precision and 0.971 recall.

The Random Tree classifier has correctly classified 534 articles out of 544 in the dataset corpus resulting in 98.16% classification accuracy. The total number of the incorrectly classified articles using the Random Tree classifier is 10 articles out of 544, which are distributed as follows: the neutral class has two incorrectly classified articles out of 134, hence, it has 132 articles true positive, four articles false positive and two articles false negative. The down class has four incorrectly classified articles out of 184, which means it has 180 articles true positive, two articles false positive and four articles false negative. The up class has four incorrectly classified articles out of 138 and so it has 134 articles true positive, four articles false positive and four articles false negative. While the critical down and the critical up classes have all the articles correctly classified resulting in zero articles false positive and zero articles false negative.

The distribution of the correctly classified articles, precision and recall for the five classes are as follows: 132 articles out of 134 for the neutral class with 0.971 precision and 0.985 recall, 180 articles out of 184 for the down class with 0.989 precision and 0.978 recall, 134 articles out of 138 for the up class with 0.971 precision and recall, 54 articles out of 54 for the critical down class and 34 articles out of 34 for the critical up class with 1.000 precision and recall for both classes.

This shows that tree classifiers (RF and Random Tree) outperform Bayes classifier (Bayes Net), rules classifier (Decision Table) and meta classifiers (Bagging and Rotation Forest) when applied on a large dataset. In addition, the run time of the tree classifiers in our experiment was less than those of the other types of classifiers. Examples of the extracted classified unigrams features used to support in the classification of the news articles in phase two experiment one is available in Appendix B.

The results yielded by applying EM clustering technique in phase two experiment one of the implementation showed that the majority of the extracted classified unigrams features belong to the economic cluster. These results led to clustering 83% of the news articles to the economic cluster, 10% of the news articles to the social cluster and 7% of the news articles to the political cluster.

Table 4.21 Phase 2-Experiment 1 the RF and ADTRee Classification Performance for Threshold > 2

Classifier	Cross Validation (Folds)	Accuracy %	Precision (Critical Down class)	Recall (Critical Down class)	Precision (Down class)	Recall (Down class)	Precision (Neutral class)	Recall (Neutral class)	Precision (Up class)	Recall (UP class)	Precision (Critical Up class)	Recall (Critical UP class)
RF	5	90.9926	1.000	1.000	0.891	0.935	0.922	0.881	0.91	0.884	0.829	0.853
	10	96.3235	1.000	1.000	0.952	0.973	0.961	0.925	0.971	0.964	0.944	1.000
	20	97.2426	1.000	1.000	0.963	0.989	0.97	0.9551	0.97	0.949	1.000	1.000
	30	97.2426	1.000	1.000	0.968	0.995	0.985	0.955	0.977	0.942	0.895	1.000
	40	***98.3456***	***1.000***	***1.000***	***0.989***	***0.989***	***0.985***	***0.985***	***0.964***	***0.971***	***1.000***	***0.971***
	50	98.1618	1.000	1.000	0.984	0.989	0.977	0.97	0.971	0.971	1.000	1.000
ADTree	5	61.0294	1.000	1.000	0.587	0.549	0.657	0.53	0.538	0.71	0.286	0.235
	10	62.6838	1.000	1.000	0.581	0.685	0.641	0.493	0.583	0.659	0.286	0.118
	20	64.1544	1.000	1.000	0.648	0.641	0.568	0.59	0.629	0.638	0.345	0.294
	30	***66.3603***	***1.000***	***1.000***	***0.654***	***0.679***	***0.625***	***0.634***	***0.61***	***0.645***	***0.471***	***0.235***
	40	64.3382	1.000	1.000	0.63	0.63	0.597	0.619	0.593	0.645	0.471	0.235
	50	65.9926	1.000	1.000	0.636	0.674	0.604	0.627	0.652	0.638	0.429	0.265

Table 4.22 Phase 2-Experiment 1 the J48 and J48graft Classification Performance for Threshold > 2

Classifier	Cross Validation (Folds)	Accuracy %	Precision (Critical Down class)	Recall (Critical Down class)	Precision (Down class)	Recall (Down class)	Precision (Neutral class)	Recall (Neutral class)	Precision (Up class)	Recall (Up class)	Precision (Critical Up class)	Recall (Critical UP class)
J48	5	77.2059	0.931	1.000	0.766	0.712	0.704	0.716	0.824	0.812	0.622	0.676
	10	80.1471	0.964	1.000	0.838	0.788	0.746	0.746	0.752	0.812	0.781	0.735
	20	83.4559	1.000	1.000	0.856	0.837	0.775	0.799	0.807	0.79	0.811	0.882
	30	***84.0074***	***0.964***	***1.000***	***0.838***	***0.842***	***0.837***	***0.843***	***0.864***	***0.783***	***0.628***	***0.794***
	40	82.9044	1.000	1.000	0.856	0.810	0.779	0.813	0.801	0.790	0.750	0.882
	50	81.4338	1.000	1.000	0.867	0.783	0.794	0.776	0.739	0.819	0.700	0.824
J48graft	5	80.1471	0.964	1.000	0.838	0.788	0.746	0.746	0.752	0.812	0.781	0.735
	10	80.5147	0.931	1.000	0.823	0.81	0.742	0.731	0.794	0.812	0.781	0.735
	20	83.0882	1.000	1.000	0.846	0.837	0.772	0.784	0.807	0.79	0.811	0.882
	30	***84.0074***	***0.964***	***1.000***	***0.838***	***0.842***	***0.837***	***0.843***	***0.864***	***0.783***	***0.628***	***0.794***
	40	82.9044	1.000	1.000	0.856	0.81	0.779	0.813	0.801	0.79	0.75	0.882
	50	81.4338	0.964	1.0000	0.867	0.783	0.794	0.776	0.748	0.819	0.7	0.824

Table 4.23 Phase 2-Experiment 1 the Decision Stump and Random Tree Classification Performance for Threshold > 2

Classifier	Cross Validation (Folds)	Accuracy %	Precision (Critical Down class)	Recall (Critical Down class)	Precision (Down class)	Recall (Down class)	Precision (Neutral class)	Recall (Neutral class)	Precision (Up class)	Recall (Up class)	Precision (Critical Up class)	Recall (Critical UP class)
Decision Stump	***5-10-20-30-40-50***	***43.75***	***1.000***	***1.000***	***0.376***	***1.000***	***0***	***0***	***0***	***0***	***0***	***0***
Random Tree	5	89.8897	1.000	1.000	0.886	0.929	0.917	0.821	0.879	0.899	0.833	0.882
	10	93.9338	1.000	1.000	0.886	0.929	0.967	0.881	0.879	0.899	0.895	1.000
	20	97.0588	1.000	1.000	0.978	0.978	0.969	0.94	0.944	0.971	1.000	1.000
	30	97.2426	1.000	1.000	0.968	0.984	0.985	0.97	0.977	0.942	0.895	1.000
	40	***98.1618***	***1.000***	***1.000***	***0.989***	***0.978***	***0.971***	***0.985***	***0.971***	***0.971***	***1.000***	***1.000***
	50	97.6103	1.000	1.000	0.989	0.984	0.985	0.955	0.964	0.971	0.895	1.000

Table 4.24 Phase 2-Experiment 1 the Bayes Net and Bagging Classification Performance for Threshold > 2

Classifier	Cross Validation (Folds)	Accuracy %	Precision (Critical Down class)	Recall (Critical Down class)	Precision (Down class)	Recall (Down class)	Precision (Neutral class)	Recall (Neutral class)	Precision (Up class)	Recall (Up class)	Precision (Critical Up class)	Recall (Critical UP class)
Bayes Net	5	66.7279	1.000	1.000	0.595	0.663	0.961	0.366	0.604	0.906	0.481	0.382
	10	68.3824	1.000	1.000	0.618	0.712	0.960	0.358	0.630	0.913	0.464	0.382
	20	68.5662	1.000	1.000	0.618	0.696	0.963	0.388	0.622	0.906	0.500	0.412
	30	69.4853	1.000	1.000	0.625	0.734	0.963	0.388	0.644	0.906	0.462	0.353
	40	70.5882	1.000	1.000	0.628	0.734	0.965	0.410	0.656	0.928	0.522	0.353
	50	***71.1397***	***1.000***	***1.000***	***0.646***	***0.745***	***0.964***	***0.403***	***0.660***	***0.942***	***0.480***	***0.353***
Bagging	5	79.7794	1.000	1.000	0.730	0.853	0.861	0.694	0.774	0.819	0.810	0.500
	10	84.5588	1.000	1.000	0.816	0.870	0.858	0.769	0.822	0.870	0.821	0.676
	20	84.375	1.000	1.000	0.788	0.908	0.917	0.746	0.811	0.841	0.846	0.647
	30	86.0294	0.964	1.000	0.814	0.902	0.889	0.776	0.850	0.862	0.862	0.735
	40	85.1103	1.000	1.000	0.809	0.897	0.890	0.784	0.830	0.848	0.815	0.647
	50	***87.6838***	***1.000***	***1.000***	***0.851***	***0.897***	***0.949***	***0.828***	***0.826***	***0.891***	***0.800***	***0.706***

Table 4.25 Phase 2-Experiment 1 the Rotation Forest and Decision Table Classification Performance for Threshold > 2

Classifier	Cross Validation (Folds)	Accuracy %	Precision (Critical Down class)	Recall (Critical Down class)	Precision (Down class)	Recall (Down class)	Precision (Neutral class)	Recall (Neutral class)	Precision (Up class)	Recall (Up class)	Precision (Critical Up class)	Recall (Critical UP class)
Rotation Forest	5	86.2132	1.000	1.000	0.822	0.902	0.858	0.769	0.867	0.848	0.879	0.853
	10	88.4191	1.000	1.000	0.901	0.886	0.854	0.828	0.846	0.877	0.941	0.914
	20	89.5221	1.000	1.000	0.921	0.891	0.881	0.881	0.857	0.87	0.816	0.912
	30	90.2574	1.000	1.000	0.914	0.918	0.862	0.888	0.887	0.855	0.912	0.912
	40	***92.8309***	***1.000***	***1.000***	***0.949***	***0.913***	***0.893***	***0.933***	***0.901***	***0.92***	***0.969***	***0.912***
	50	92.2794	1.000	1.000	0.94	0.929	0.904	0.918	0.891	0.891	0.912	0.912
Decision Table	5	64.8897	1.000	0.944	0.593	0.728	0.577	0.612	0.683	0.514	0.714	0.441
	10	67.0956	1.000	0.926	0.616	0.723	0.583	0.657	0.752	0.594	0.667	0.353
	20	***70.7721***	***1.000***	***0.907***	***0.701***	***0.712***	***0.576***	***0.731***	***0.775***	***0.674***	***0.778***	***0.412***
	30	70.0368	1.000	0.981	0.638	0.739	0.633	0.709	0.769	0.601	0.7	0.412
	40	68.9338	1.000	0.944	0.639	0.701	0.614	0.724	0.757	0.609	0.636	0.412
	50	70.2206	1.000	0.926	0.684	0.766	0.59	0.709	0.771	0.609	0.667	0.353

4.5.3 Results of Experiment Two of Phase Two

The above sets of results based on the five classes have enhanced the classification performance of SMRF-TM. The next task is to assess the effect of using bigrams rather than unigrams based features on the performance of SMRF-TM. Tables 4.3, 4.4, 4.5, 4.6, 4.7, 4.8, 4.9, 4.10 and 4.11 show that the best performance is achieved with a threshold > 2 for unigrams based features and threshold >1 for bigrams based features. This indicates that the bigrams require a larger dataset than the unigrams in order to enhance the performance of SMRF-TM.

The implementation results of experiment two of phase two are summarised in Tables 4.26, 4.27, 4.28 and 4.29. The RF classifier has correctly classified 538 articles out of 544 in the dataset corpus resulting in 98.89% classification accuracy. The total number of the incorrectly classified articles using the RF classifier is six articles out of 544, which are distributed as follows: the neutral class has two incorrectly classified articles out of 134, so it has 132 articles true positive, four articles false positive and two articles false negative. The down class has two incorrectly classified articles out of 184, which means it has 182 articles true positive, zero articles false positive and two articles false negative. The up class has two incorrectly classified articles out of 138, 136 articles true positive, two articles false positive and two articles false negative. Regarding the critical down and the critical up classes they did not have any misclassified articles, hence, they have zero articles false positive and zero articles false negative.

Consequently, the distribution of the correctly classified articles, precision and recall for the five classes are as follows: the neutral class has 132 articles out of 134 correctly classified with 0.971 precision and 0.985 recall. The down class has 182 articles out of 184 correctly classified with 1.000 precision and 0.989 recall. The up class has 136 articles out of 138 correctly classified with 0.986 precision and recall. While the critical down and critical up classes they have all the articles correctly classified and so they have 1.000 precision and recall for both classes.

Table 4.26 Phase 2-Experiment 2 the RF and ADTRee Classification Performance for Threshold > 1

Classifier	Cross Validation (Folds)	Accuracy %	Precision (Critical Down class)	Recall (Critical Down class)	Precision (Down class)	Recall (Down class)	Precision (Neutral class)	Recall (Neutral class)	Precision (Up class)	Recall (UP class)	Precision (Critical Up class)	Recall (Critical UP class)
RF	5	91.5441	1.000	1.000	0.929	0.929	0.901	0.881	0.874	0.906	0.938	0.882
	10	95.4044	1.000	1.000	0.957	0.973	0.946	0.910	0.943	0.957	0.941	0.941
	20	97.4265	1.000	1.000	0.978	0.989	0.955	0.948	0.971	0.964	1.000	1.000
	30	95.7721	1.000	1.000	0.953	0.984	0.955	0.940	0.962	0.928	0.914	0.941
	40	***98.8971***	***1.000***	***1.000***	***1.000***	***0.989***	***0.971***	***0.985***	***0.986***	***0.986***	***1.000***	***1.000***
	50	***98.8971***	***1.000***	***1.000***	***0.989***	***1.000***	***0.985***	***0.970***	***0.986***	***0.986***	***1.000***	***1.000***
LADTree	5	66.3603	1.000	1.000	0.574	0.761	0.857	0.448	0.627	0.754	0.3	0.088
	10	66.7279	0.649	1.000	0.588	0.766	0.786	0.493	0.649	0.71	0.308	0.118
	20	68.3824	1.000	1.000	0.606	0.777	0.833	0.485	0.653	0.79	0.111	0.029
	30	67.4632	1.000	1.000	0.589	0.755	0.835	0.493	0.655	0.783	0.000	0.000
	40	68.75	1.000	1.000	0.606	0.777	0.87	0.5	0.651	0.797	0.000	0.000
	50	***69.8529***	***1.000***	***1.000***	***0.611***	***0.777***	***0.864***	***0.522***	***0.677***	***0.819***	***0.000***	***0.000***

Table 4.27 Phase 2-Experiment 2 the Decision Stump, Random Tree and J48 Classification Performance for Threshold > 1

Classifier	Cross Validation (Folds)	Accuracy %	Precision (Critical Down class)	Recall (Critical Down class)	Precision (Down class)	Recall (Down class)	Precision (Neutral class)	Recall (Neutral class)	Precision (Up class)	Recall (Up class)	Precision (Critical Up class)	Recall (Critical UP class)
Decision Stump	***5-10-20-30-40-50***	***43.75***	***0.964***	***1.000***	***0.377***	***1.000***	***0.000***	***0.000***	***0.000***	***0.000***	***0.000***	***0.000***
Random Tree	5	88.2353	1.000	1.000	0.879	0.870	0.902	0.821	0.818	0.913	0.938	0.882
	10	94.4853	1.000	1.000	0.957	0.967	0.983	0.851	0.905	0.971	0.850	1.000
	20	97.7941	1.000	1.000	0.989	0.989	0.970	0.955	0.957	0.971	1.000	1.000
	30	***98.5294***	***1.000***	***1.000***	***0.989***	***1.000***	***0.985***	***0.970***	***0.985***	***0.971***	***0.944***	***1.000***
	40	98.3456	1.000	1.000	0.989	0.995	0.985	0.963	0.965	0.986	1.000	0.971
	50	96.5074	1.000	1.000	0.973	0.984	0.962	0.940	0.949	0.949	0.943	0.971
J48	5	77.5735	0.844	1.000	0.779	0.804	0.802	0.664	0.750	0.761	0.667	0.765
	10	77.7574	0.818	1.000	0.803	0.799	0.802	0.664	0.752	0.790	0.615	0.706
	20	79.2279	0.900	1.000	0.788	0.810	0.798	0.709	0.750	0.783	0.781	0.735
	30	80.5147	0.900	1.000	0.823	0.810	0.817	0.731	0.747	0.790	0.757	0.824
	40	81.0662	0.964	1.000	0.825	0.793	0.820	0.746	0.750	0.826	0.730	0.794
	50	***81.25***	***0.964***	***1.000***	***0.841***	***0.804***	***0.811***	***0.739***	***0.750***	***0.826***	***0.711***	***0.794***

Table 4.28 Phase 2-Experiment 2 the Bayes Net and Bagging Classification Performance for Threshold > 1

Classifier	Cross Validation (Folds)	Accuracy %	Precision (Critical Down class)	Recall (Critical Down class)	Precision (Down class)	Recall (Down class)	Precision (Neutral class)	Recall (Neutral class)	Precision (Up class)	Recall (Up class)	Precision (Critical Up class)	Recall (Critical UP class)
Bayes Net	5	71.875	1.000	1.000	0.580	0.908	0.962	0.560	0.771	0.659	0.667	0.118
	10	73.8971	1.000	1.000	0.608	0.918	0.963	0.582	0.776	0.703	0.667	0.118
	20	73.5294	1.000	1.000	0.607	0.897	0.963	0.575	0.758	0.725	0.667	0.118
	30	74.8162	1.000	1.000	0.624	0.902	0.974	0.567	0.764	0.775	0.667	0.118
	40	74.2647	1.000	1.000	0.622	0.886	0.963	0.59	0.743	0.754	0.667	0.118
	50	***75.3676***	***1.000***	***1.000***	***0.630***	***0.908***	***0.975***	***0.590***	***0.768***	***0.768***	***0.667***	***0.118***
Bagging	5	78.3088	0.964	1.000	0.700	0.875	0.888	0.649	0.775	0.797	0.778	0.412
	10	79.5956	0.964	1.000	0.719	0.875	0.897	0.649	0.796	0.819	0.720	0.529
	20	80.5147	0.982	1.000	0.740	0.864	0.875	0.679	0.786	0.826	0.800	0.588
	30	***81.9853***	***0.982***	***1.000***	***0.772***	***0.864***	***0.870***	***0.746***	***0.799***	***0.804***	***0.759***	***0.647***
	40	81.25	0.982	1.000	0.743	0.880	0.921	0.694	0.776	0.826	0.826	0.559
	50	81.8015	0.964	1.000	0.765	0.886	0.888	0.709	0.799	0.833	0.750	0.529

Table 4.29 Phase 2-Experiment 2 the Rotation Forest and Decision Table Classification Performance for Threshold > 1

Classifier	Cross Validation (Folds)	Accuracy %	Precision (Critical Down class)	Recall (Critical Down class)	Precision (Down class)	Recall (Down class)	Precision (Neutral class)	Recall (Neutral class)	Precision (Up class)	Recall (Up class)	Precision (Critical Up class)	Recall (Critical UP class)
Rotation Forest	5	84.1912	1.000	1.000	0.832	0.864	0.822	0.791	0.82	0.826	0.806	0.735
	10	85.8456	1.000	1.000	0.874	0.87	0.859	0.821	0.817	0.841	0.73	0.794
	20	85.4779	1.000	1.000	0.854	0.859	0.866	0.866	0.813	0.819	0.75	0.706
	30	***86.2132***	***1.000***	***1.000***	***0.849***	***0.886***	***0.877***	***0.851***	***0.819***	***0.819***	***0.833***	***0.735***
	40	***86.2132***	***1.000***	***1.000***	***0.861***	***0.875***	***0.869***	***0.843***	***0.807***	***0.848***	***0.857***	***0.706***
	50	85.2941	1.000	1.000	0.865	0.87	0.857	0.806	0.799	0.833	0.771	0.794
Decision Table	5	69.6691	1.000	0.944	0.597	0.788	0.798	0.619	0.707	0.630	0.565	0.382
	10	69.4853	1.000	0.926	0.632	0.783	0.648	0.590	0.714	0.616	0.800	0.588
	20	71.1397	1.000	0.963	0.632	0.793	0.664	0.619	0.775	0.623	0.800	0.588
	30	72.4265	1.000	0.963	0.635	0.793	0.730	0.627	0.740	0.659	0.875	0.618
	40	72.6103	1.000	0.944	0.655	0.793	0.695	0.664	0.761	0.645	0.800	0.588
	50	***73.5294***	***1.000***	***0.907***	***0.651***	***0.832***	***0.773***	***0.634***	***0.764***	***0.681***	***0.704***	***0.559***

The Random Tree classifier has correctly classified 536 articles out of 544 in the dataset corpus resulting in 98.52% classification accuracy. The total number of the incorrectly classified articles using the Random Tree classifier is eight articles out of 544, which are distributed as follows: the neutral class has six incorrectly classified articles out of 134, which means that it has 130 articles true positive, two articles false positive and four articles false negative. The up class has 134 articles true positive, two articles false positive and four articles false negative. The down class has 184 articles true positive and two articles false positive. The critical up class has 34 articles true positive and two articles false positive. Regarding the critical down class it did not has any misclassified articles, so it has 54 articles true positive, zero articles false positive and zero articles false negative.

The distribution of the correctly classified articles, precision and recall for the five classes are as follows: the neutral class has 130 articles correctly classified out of 132, which should be 134 with 0.985 precision and 0.97 recall. The up class has 134 articles correctly classified out of 136, which should be 138 with 0.985 precision and 0.971 recall. The down class has 184 articles correctly classified out of 186, which should be 184 with 0.989 precision and 1.000 recall. The critical up class has 34 articles correctly classified out of 36, which should be 34 with 0.944 precision and 1.000 recall. As for the critical down class it has all the articles correctly classifies, which means that it has 1.000 precision and recall.

These results demonstrate that the tree classifiers (RF and Random Tree) perform much better than the Bayes classifier (Bayes Net), rules classifier (Decision Table) and the meta classifiers (Bagging and Rotation Forest) when applied on large datasets. They also show that bigrams based features/tokens in experiment two of phase two can enhance the classification performance of SMRF-TM. Examples of the classified bigrams features used to support the classification of the news articles in experiment two of phase two is available in Appendix C.

The results yielded by applying EM clustering technique in phase two experiment two of the implementation showed that majority of the extracted classified bigrams features belong to the economic cluster. Hence, 83% of the news articles were clustered as economic, 10% clustered as social and 7% clustered as political.

4.6 Summary and Conclusions

The classification results of phase one in Tables 4.15, 4.16 and 4.17, using the extracted features with threshold >2 show that the tree classifiers (ADTree and RF) outperformed the other types of tested classifiers, which are meta classifiers (e.g. Bagging), rules classifier (e.g. Decision Table) and Bayes classifier (e.g. Bayes Net). This indicates that the tree classifiers were more capable than the other types of classifiers in retrieving hidden information and the important relations between the extracted features, which supported them to have a better performance. These results support the first hypothesis of this research, which states that the application of the RF to the domain of SM textual data can elicit the crucial relationships between the extracted features leading to the enhancement of the classification performance, and, therefore, it can be an effective predictive measure for the stock market movements. The application of the random forest on the stock market domain is considered as the major contribution of this research.

In phase two experiment one, the 544 articles in the dataset corpus were used to check the effect of expanding the dataset on the performance of SMRF-TM approach. Classification classes are extended to five classes (critical down, down, neutral, up and critical up) rather than the two classes used in phase one. The experimentation results proved that by doing so the classification performance for the extracted features and the news articles was enhanced especially regarding the random forest classifier. This supported the hypothesis of this research, which is that by extending the classes from three to five in SMRF-TM, the classification performance of both the extracted features as

well as the news articles is enhanced. Consequently, this extension also contributes to the novel approach of SMRF-TM.

In addition, the experimentation results of phase two showed that the tree classifiers (RF and Random Tree) perform much better than the Bayes classifier (Bayes Net), rules classifier (Decision Table) and the meta classifiers (Bagging and Rotation Forest) when applied on large datasets. While the experimentation results of phase two experiment two showed that bigrams based features/tokens enhanced the classification performance of SMRF-TM compared to the outcome achieved in phase one.

This research had a limitation in the application of the Expectation Maximisation (EM) clustering technique, which is in the use of Financial Times as the only source for our dataset. Since the Financial Times only publishes financial news, we had the majority of the features belonging to the economic cluster. However, the application of EM clustering technique can also be regarded a novel contribution as SMRF-TM is able not only to classify the features/articles according to the predicted influence they have on Dubai's SM movements, but also able to examine the reasons behind any movement in Dubai's SM.

In the following chapter, the validation and evaluation methods applied to validate the SMRF-TM approach and to evaluate the classification performance and the results achieved from the two phases of the implementation are explained. It also discusses the strengths and the limitations of the validation and evaluation processes.

Chapter 5 Validation and Evaluation Methods Used in the SMRF-TM Approach

5.1 Introduction

In the previous chapter the architecture of the proposed Stock Market Random Forest-Text Mining (SMRF-TM) approach based on the finding of the literature review and the Text Mining (TM) theoretical basis was described in addition to the development stages of the SMRF-TM approach. Chapter four also described the implementation phases and presented the results from each phase. This chapter explains the validation and evaluation methods applied to validate and evaluate both the SMRF-TM approach and the results achieved from the two phases of the implementation. It also summarises the strengths and the limitations of the validation and evaluation processes.

5.2 Validation Approach

In this section the different validation methods, which are used in order to validate both the SMRF-TM approach and the results yielded by the two phases of the implementation are described.

The validation approach applied in this research is a qualitative approach because a continuous validation process was required in this research for the following reasons. At the early stages, we needed the opinions of domain experts about the data source to be used for retrieving the required dataset for the purpose of the analysis of Dubai Debt stand still 2009 in order to make sure that it is a reliable source. Also, we needed the experts' opinions within the period of designing the SMRF-TM approach to ensure that we have encountered most of the significant tasks, which can support the analysis of Dubai's stock market while implementing the SMRF-TM approach. Last but not least, the qualitative validation of the extracted features against the domain experts' opinions was

very important in this research to check wither the extracted features and their relationships are appropriate for the analysis of Dubai's stock market domain or not.

The quantitative approach is also used in this research through the usage of cross validation in order to examine the learning capabilities of the SMRF-TM approach specially the Random Forest (RF) classifier. In addition, we used some quantitative evaluation measures such as precision, recall, F-measure to evaluate the classification performance of random forest in classifying the extracted features and the news articles according to their influence on the Stock Market (SM) movements. Consequently, the quantitative validation process used did not only support us to ensure that the classified features are good indicators, which can affect Dubai's SM movements, but also supported us in measuring the classification performance of RF in classifying the extracted features and the news articles.

The qualitative and quantitative validation methods, which are used in this research, are described in detail in the following sections.

5.3 SMRF-TM Validation Using a Quantitative Approach

5.3.1 K-Fold Cross Validation

This section discusses k-fold cross validation, which was the quantitative validation method used to examine the learning capabilities of random forest classifier used in this research.

Larson (1931) realised that training an algorithm and evaluating its performance using the same dataset produces overoptimistic results. Later, Geisser (1975), Stone (1974) and Mosteller and Tukey (1968) stated that testing the output of an algorithm on a new dataset would result in a better estimate of its performance. Hence, the idea of splitting the original dataset available for research into a training set to train the algorithm and a validation test set to evaluate the performance of the algorithm was raised because the

availability of data for some applications are limited (Wong 2017, Wong 2015, Arlot & Celisse 2010). This idea is the main strategy adopted by cross validation, which led cross validation to be a popular technique used in academic and commercial statistical packages to evaluate the performance of a classifier when no separate test set is available. Consequently, cross validation can give an indication of how well the learner classifier will do when it is asked to classify a new dataset. The k-fold cross validation technique is considered the default technique among the different cross validation existing techniques such as Monte-Carlo, Leave One Out, double cross validation, etc. (Wong 2017, Triba *et al.* 2015, Wong 2015, Moreno-Torres *et al.* 2012, Rodriguez *et al.* 2010, Arlot & Celisse 2010).

Recently, Donate *et al.* (2013) deployed k-fold cross validation in order to get the average forecast from different forecasts produced by multiple models, which are trained on diverse data subsets. In this research, we used the k-fold cross validation in the SMRF-TM approach to measure the learning capabilities of the random forest classifier as explained previously in chapter 4, where the original dataset sample is randomly partitioned into k disjoint and approximately equal size subsamples. Then k-1 out of the k subsamples are used as a training dataset to train the random forest and the remaining one subsample is used as a validation dataset to evaluate the classification performance of random forest. The cross validation process is repeated k times where each of the k subsamples is used only once as the validation dataset. To produce a single estimation the k results from the folds are averaged. The accuracy measured by k-fold cross validation for a dataset collected from a certain source represents the probability of correctly predicting the class value of a new instance coming from the same source (Wong 2017, Barrow & Crone 2016, Wong 2015).

One of the advantages of using k-fold cross validation method in the SMRF-TM approach is that all observations are used for both training and validation while each observation is used for validation only once, which makes the results more reliable (Wong

2017, Jiang & Chen 2016). In addition, using k-fold cross validation supported us not to lose significant modelling or testing capability, which would have been lost if we had partitioned the available dataset into a separate training and test sets since the available data related to Dubai debt standstill 2009 in the Financial Times was not big enough to be partitioned without encountering such loss. The k-fold cross validation avoids over fitting the model with statistically insignificant variables and produces more powerful models (Wong 2017, Jiang & Chen 2016, Boxell 2015).

In this research, different values for (k) in the k-fold cross validation were used (5, 10, 20, 30, 40 and 50) in order to compare performance and check which value of (k) yields the best classification performance as shown previously in chapter 4. Using the k-fold cross validation supported in the calculation of the precision, recall, F-measure and the confusion matrix for all the classification algorithms tested in classifying the extracted features and the news articles, which were of a great support for the purpose of the analysis of Dubai stock market.

5.3.2 K-Fold Cross Validation Results

The results yielded by applying the K-fold cross validation are used to determine the performance of the tested classifiers in learning from the training dataset, which were discussed in detail in chapter 4. The usage of different values for (k) in the k-fold cross validation was to check which value of (k) enhances the learning capabilities of the RF classifier in order to achieve the best classification performance. As shown in chapter 4, RF achieved the best classification performance in phase one 88.82% while using 30-folds cross validation, in phase two experiment one 98.35% while using 40-folds cross validation and in phase two experiment two 98.89% while using 40-folds cross validation.

Accuracy, precision, recall, F-measure and confusion matrix, which are applied to determine the classification performance of the classifier algorithms used in the SMRF-TM approach are considered to be the most commonly utilised measures in information

systems and machine learning (Kumar & Khatri 2017, Deng *et al.* 2016, Ziółko 2015, Ponciano *et al.* 2015, Estrada & Jepson 2009, Ziółko *et al.* 2007, Grocholewski 1997, Rijsbergen 1979).

i) Accuracy

Accuracy of a classifier refers to the closeness of a measured value to the true value and it indicates the capability of the classifier to successfully classify new data values. Measuring the classification accuracy of classifier algorithms is a very important task in order to determine the classifiers' performance (Kumar & Khatri 2017, Deng *et al.* 2016). Accuracy can be defined as the proportion of the total number of classifications, which were correct (Deng *et al.* 2016).

The experimentation results discussed in chapter 4 have shown that the highest classification accuracies in phase one were achieved by ADTree 91.9255% and RF 88.8199%. In phase two experiment one the highest classification accuracies were achieved by RF 98.3456% and Random Tree 98.1618%, while in phase two experiment two RF has achieved 98.8971% classification accuracy followed by Random Tree, which has achieved 98.5294% classification accuracy.

ii) Precision

Precision is regarded as one of the common measures used in order to determine the classification performance but it is not the same as accuracy; precision refers to the closeness of two or more measurements to each other, while accuracy, refers to the closeness of a measured value to the true value. So being precise does not necessarily mean being accurate and vice versa. Hence, each of precision and accuracy has its own definite meaning (Kumar & Khatri 2017, Deng *et al.* 2016, Ziółko 2015, Ponciano *et al.* 2015, Streiner & Norman 2006).

Even though precision cannot be used as a synonym for reliability, it is still considered as one of the main components of reliability and that is why it is treated as an important validation measure, which is used in our research (Kumar & Khatri 2017,

Ziółko 2015, Ponciano *et al.* 2015, Streiner & Norman 2006, Goodwin & Leech 2003, Streiner 2003).

In this research, the value of the precision measure illustrates the closeness of the classification results of the tested classifiers per each class of the five classes for the total number of runs undertaken while applying the k-fold cross validation. Precision is achieved through specifying the truly positive instances in their proportion to the totally predicted positive instances (Kumar & Khatri 2017, Deng *et al.* 2016). The equation used in this research to calculate the precision of the classification results yielded by the tested classifiers for each of the five classes deployed in the implementation was explained in chapter 4.

The experimentation results discussed in chapter 4 have shown that in phase one the ADTree, which achieved the highest classification accuracy, had precision 0.917 for the down class and a slightly higher precision 0.925 for the up class. The RF, which achieved the second highest classification accuracy, had precision 0.887 for the down class and a bit higher precision 0.892 for the up class.

In phase two experiment one the RF, which achieved the highest classification accuracy, had precision 1.000 for the critical down class, 0.989 for the down class, 0.985 for the neutral class, 0.964 for the up class and 1.000 for the critical up class. The Random Tree, which achieved the second highest classification accuracy, had precision 1.000 for the critical down class, 0.989 for the down class, 0.971 for the neutral class, 0.971 for the up class and 1.000 for the critical up class.

In phase two experiment two RF has achieved the highest classification accuracy with precision 1.000 for the critical down class, 1.000 for the down class, 0.971 for the neutral class, 0.986 for the up class and 1.000 for the critical up class. Followed by Random Tree, which had precision 1.000 for the critical down class, 1.000 for the down class, 0.985 for the neutral class, 0.985 for the up class and 1.000 for the critical up class.

iii) Recall

Recall is the ability to remember (bring back) things from the past (memory). In this research, it refers to the ability of the tested classifiers to remember what it learned from the training dataset to apply it on the new testing dataset. So, the value of the recall measure is a very good indicator for the classifiers' learning capabilities. Recall represents the proportion of real positive values, which are really correctly positive (Kumar & Khatri 2017, Deng *et al.* 2016, Ziółko 2015, Ponciano *et al.* 2015, Estrada & Jepson 2009). The equation used in this research to calculate the recall of the classification results yielded by the RF and the other tested classifiers for each of the five classes used in the implementation was given in chapter 4.

As shown by the experimentation results discussed in chapter 4, ADTree, which achieved the highest classification accuracy in phase one, had recall 0.974 for the down class and 0.787 for the up class. The RF, which achieved the second highest classification accuracy, had recall 0.965 for the down class and 0.702 for the up class.

In phase two experiment one the RF, which achieved the highest classification accuracy, had recall 1.000 for the critical down class, 0.989 for the down class, 0.985 for the neutral class, 0.971 for the up class and 0.971 for the critical up class. The Random Tree, which achieved the second highest classification accuracy, had recall 1.000 for the critical down class, 0.978 for the down class, 0.985 for the neutral class, 0.971 for the up class and 1.000 for the critical up class.

In phase two experiment two RF has achieved the highest classification accuracy with recall 1.000 for the critical down class, 0.989 for the down class, 0.985 for the neutral class, 0.986 for the up class and 1.000 for the critical up class. Followed by Random Tree, which had recall 1.000 for the critical down class, 1.000 for the down class, 0.970 for the neutral class, 0.971 for the up class and 1.000 for the critical up class.

iv) F-measure

The F-measure is a combined metric, which represents a balanced harmonic mean of precision and recall metrics and it is sometimes referred to as effectiveness measure. The F-measure is considered as a standard performance index commonly used in machine learning to evaluate the classification performance in precision and recall space (Kumar & Khatri 2017, Guo *et al.* 2016, Deng *et al.* 2016, Maratea *et al.* 2014, Han *et al.* 2011, Lazarevic-McManus *et al.* 2008).

There are some differences in the classification abilities of a classifier to different classes in multi-class classification, which are not easy to be reflected using any single performance index. The F-measure is capable of reflecting such differences because through the combination of precision and recall it holds all the information included in a confusion matrix, which is explained in the following section (Deng *et al.* 2016).

The F-measure methods are able to evaluate the tested algorithms and produce an objective comparison of two or more algorithms and that is why we used it in our comparative study of the tested classifiers as explained previously in chapter 4 (Lazarevic-McManus *et al.* 2008). The equation used in this research to calculate the F-measure of the classification results yielded by the tested classifiers for each of the five classes is shown in chapter 4.

As shown by the experimentation results discussed in chapter 4, ADTree, which achieved the highest classification accuracy in phase one, had F-measure 0.945 for the down class and 0.850 for the up class. The RF, which achieved the second highest classification accuracy, had F-measure 0.924 for the down class and 0.786 for the up class.

In phase two experiment one the RF, which achieved the highest classification accuracy, had F-measure 1.000 for the critical down class, 0.989 for the down class, 0.985 for the neutral class, 0.967 for the up class and 0.985 for the critical up class. The Random Tree, which achieved the second highest classification accuracy, had F-

measure 1.000 for the critical down class, 0.983 for the down class, 0.978 for the neutral class, 0.971 for the up class and 1.000 for the critical up class.

In phase two experiment two RF has achieved the highest classification accuracy with F-measure 1.000 for the critical down class, 0.994 for the down class, 0.978 for the neutral class, 0.986 for the up class and 1.000 for the critical up class. Followed by Random Tree, which had F-measure 1.000 for the critical down class, 1.000 for the down class, 0.977 for the neutral class, 0.978 for the up class and 1.000 for the critical up class.

v) Confusion Matrix

Confusion matrix is a 2-D matrix considered as a common validation measure in machine learning, which holds information about predicted and actual classifications done by a classifier (classification model) and it is usually used to evaluate the performance of the classifier using the data in the matrix (Deng *et al.* 2016, Sammut & Webb 2011, Kohavi & Provost 1998). One dimension (columns) represents the actual class of an instance, while the other dimension (rows) represents the predicted class for that instance. Figure 5.1 shows the standard shape of a confusion matrix for a multi-class (five classes) classification, which is deployed in the implementation of the SMRF-TM approach. The classes are denoted as A_1, A_2…and A_n, while $N_{i,j}$ indicates the number of instances actually belonging to class A_i and classified as class A_j and the diagonal cells where i=j contains the number of correctly classified instances. The confusion matrix yields information in a comprehendible form. Consequently, utilising such information is considered to give great support in supervised machine learning (Deng *et al.* 2016).

		Predicted Classes			
		A_1	A_2	.....	A_n
Actual Classes	A_1	N_{11}	N_{12}	.....	N_{1n}
	A_2	N_{21}	N_{22}	.....	N_{2n}
	⋮	⋮	⋮	⋮	⋮
	A_n	N_{n1}	N_{n2}	.....	N_{nn}

Figure 5.1 Standard shape of confusion matrix for multi-class classification

In this research, the strength of the confusion matrix is that it identified the nature of the RF classification errors, in addition to their quantities. It also aided in measuring the classification performance of the RF by using the data in the matrix. An example of one of the confusion matrices produced by the SMRF-TM approach is shown below in Figure 5.2.

```
=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
 110  16   8   0   0 |   a = 0
   9 170   5   0   0 |   b = 1
   4   6 126   0   2 |   c = 2
   0   0   0  54   0 |   d = 3
   0   3   2   0  29 |   e = 4
```

Figure 5.2 Example of a confusion matrix produced by SMRF-TM

5.4 SMRF-TM Validation Using a Qualitative Approach

This section discusses the aim of the features validation and the process used to qualitatively validate the results yielded by the SMRF-TM approach.

5.4.1 The Aim of the Features Validation

The main aim of the features validation is to check whether the extracted features have an appropriate discriminative power or not. By discriminative power we mean the uniqueness or degeneracy of the extracted features, which can be quantitatively measured by the features' relevance to classification in addition to its generalisability for classification (Dehmer *et al.* 2012, Fan *et al.* 2005). The correlation between the extracted features and the corresponding class label in the training dataset can be used to measure its relevance to classification (Fan *et al.* 2005). This discriminative power is very important in supporting the SMRF-TM approach to be able to correctly classify the extracted features and the news articles according to the five classes described in the previous chapter. Validating the extracted features also ensures that the extracted features are good indicators for SM movements and that they can be used for further prediction of stock market abnormal movements.

5.4.2 Design of the Features Validation Process

In the SMRF-TM approach the features validation process adopted is a qualitative validation process, which was a continuous process since the experimentation results of the SMRF-TM approach needed to be qualitatively validated against stock market domain experts' opinion during the whole duration of the experimental work at all of the different stages. It was important to have a domain expert in order to be able to judge if the extracted features are really important and if they can affect the SM movements or not.

Consequently, some procedures had to be carried out in order to be able to apply such a qualitative validation. The first procedure was to find the appropriate experts for the study domain of this research. This required a thorough investigation/search in some of

the different fields related to the study domain of our research, which are the stock market, academic and banking fields; these fields have different and important backgrounds. The investigation/search was followed by some interviews/meetings with the most appropriate experts. The criteria used were good domain knowledge and experience as well as geographically accessible and the interviews were used to check for willingness to participate and availability. This procedure successfully identified one academic expert in finance (an assistant professor in the Arab Academy for Science and Technology, Faculty of Business Administration), one banking expert (a stock markets' Analyst in a multinational bank) and two stock market experts (one is managing director of a multinational brokering company and the other is Stocks' Analyst in a multinational brokering company). These participants were selected as they had the relevant domain knowledge and experience in addition to the availability to provide continuous, reasonable and valuable feedbacks. As mentioned above, the qualitative validation in this research was a continuous process so it started in the early stages of the research by validating all the available data sources with the experts to make sure that a reliable data source for the data collection/retrieval process was chosen. This led to the choice of the Financial Times as the research data source. Then, as soon as the implementation of the SMRF-TM approach started, a series of 28 meetings were designed to be carried out with all the experts; the meetings were to be held at their job sites after each stage of the three stages of the implementation and by the end of each experiment.

5.4.3 Implementation of the Features Validation

During the implementation of SMRF it was found that the banking expert opinions was not as beneficial as expected because he is only concerned with the stocks related to the business of the bank he is working at, but not the whole market. Hence, we relied more on the academic and the stock market experts' opinions; thus, the number of meetings was reduced to 23 meetings, which were held at their job sites.

The experimental works were carried out in two different phases, as explained previously in chapter four. Phase one was carried out through a proof of concept to examine the feasibility of our proposed approach; in this thesis, it is referred as experiment one and was based on using unigram tokens and a subset of the data retrieved while phase two was carried out through two further experiments using the complete dataset retrieved in order to compare the performance of using unigram tokens against bigram tokens.

As shown previously in chapter four, using bigrams tokens to analyse stock markets requires a larger dataset than using unigrams in order to enhance the performance of SMRF-TM. The experimentation results demonstrated that tree classifiers (RF and Random Tree) outperformed Bayes classifier (Bayes Net), rules classifier (Decision Table) and meta classifiers (Bagging and Rotation Forest) when applied on large datasets. The experimentation results of phase two experiment two specifically, showed that bigrams based features/tokens enhanced the classification performance of SMRF-TM compared to the results achieved in phase two experiment one.

The results of each of the three experiments were validated against the experts' opinions in relation to the extracted features. This continuous feedback after each experiment helped improve the MATLAB code following the recommendations of our experts. Tables 5.1 and 5.2 show examples of the unigrams and bigrams critical factors respectively, which were extracted using SMRF-TM and approved by the experts as significant in the analysis of stock markets.

Table 5.1 Examples of unigrams critical factors extracted by SMRF-TM

Beta	Margin	Volume	Offset
Broker	Order	Yield	Commodities
Dividend	Portfolio	Agent	Debentures
Exchange	Quote	Securities	Delta
Execution	Rally	Offer	Derivatives
Hedge	Sector	Assets	Diversification
Index	Spread	Bid	Equity
Inflation	Volatility	Bonds	Risk

Table 5.2 Examples of bigrams critical factors extracted by SMRF-TM

Oversell	Blue chip	Open price	Basis point
Overbuy	Chip stocks	Close price	Clearing day
Averaging down	Moving average	Annual report	Capital trust
Bear market	Margin account	Anonymous trading	Low price
Bull market	Improving market	Capital loss	Last sale
Initial public	Growth stock	Capital gain	Sale price
Public offering	Downtick	Booked orders	Issue status
Day trading	Defensive stock	Bid size	Board lot

5.4.4 Conclusions from the Features Validation

The meetings held with the academic and the SM experts were very supportive in highlighting significant tasks and challenges, which can affect the process of analysing

the stock market. For example, the handling of negative expressions such as no, not, n't, neither and nor, which was recommended by the experts, were incorporated in stage one of the SMRF-TM implementation. In addition, the consideration of bigram tokens was also included in phase two. Although the experts' recommendations have improved the results yielded by the SMRF-TM approach, one limitation was identified. This limitation concerns the reliance on Financial Times as the sole data source to train the classifiers. It was recommended, in future work, to employ a variety of data sources to train the RF in order to extract the hidden information and relationships, which capture different interpretations of the stock market news. One positive outcome for the current implementation is the financial offer of support from the managing director expert of a multinational brokering company, to extend SMRF-TM dataset to include the company's subscriptions to different news sources such as Bloomberg financial news Reuters, Wall Street Journal, and Economic Times.

5.5 Summary and Conclusions

As explained above the quantitative and qualitative validation approaches are not mutually exclusive. Consequently, in this research we deployed the quantitative and qualitative validation approaches in order to make use of the benefits of both approaches.

The use of the quantitative cross-validation approach was supportive in measuring the learning capabilities of the RF classifier. It was also very effective in the evaluation of classification performance through the calculation of a set of complementary measures such as, accuracy, precision, recall, F-measure and producing the confusion matrix, which are considered to be the most commonly utilised measures in information systems and machine learning.

The use of the qualitative validation approach had a significant support in enhancing the performance of the SMRF-TM approach by continuously checking the obtained results against the experts' opinions with differing backgrounds: academic and business/financial.

The validation of SMRF-TM experienced some challenges. Regarding the quantitative validation process, we faced a challenge, which was related to the software used to apply the RF classifier. WEKA was the software used to apply the RF classifier, and WEKA has a limited memory allocation for each run of the different classifiers tested. This caused a problem during the experimentations, particularly in the experiments of phase two, which applied the complete dataset. The use of a big dataset (i.e. 544 news articles) caused WEKA to run out of memory while applying some of the tested classifiers in phase two of the experimentation works such as ADTree, j48graft and Rotation Forest. In order to overcome this challenge a bigger memory size had to be allocated manually before each run.

Regarding the qualitative validation process, we faced another challenge, related to the inability of the experts to supply us with a reference list of critical indicators (features) related to the SM movements. Consequently, the qualitative validation process had to be an interactive and lengthy process in order to make sure that the extracted features are critical indicators, which may have profound influence on the SM movements. Through these 23 meetings we were able to ensure that the retrieved results were consistent with the experts' opinions.

The meetings held with the academic and the SM experts were very supportive in highlighting significant tasks and challenges, which can affect the process of analysing the SM. For example, the task of the negative expressions such as no, not, n't, neither and nor, which were recommended by the experts, were incorporated in stage one of the SMRF-TM implementation. In addition, the consideration of bigram tokens was also included in phase two. Although the experts' recommendations have improved the results yielded by the SMRF-TM approach, one limitation was identified, which could recognisably enhance the performance of SMRF-TM. This limitation concerns the reliance on Financial Times as the sole data source to train the classifiers. It was recommended, in future work, to employ a variety of data sources to train the RF in order to extract the

hidden information and relationships, which capture different interpretations of the stock market news. One positive outcome for the current implementation is the offer of financial support from the managing director expert of a multinational brokering company, to extend SMRF-TM dataset to include the company's subscriptions to different news sources such as Bloomberg financial news Reuters, Wall Street Journal, and Economic Times.

Chapter 6 Conclusion and Future Work

6.1 Relevance of Text Mining to the Understanding of Stock Market Movements

Knowledge discovery is a fast-growing field of research providing hidden and valuable knowledge stored in ever increasing amounts of data. We have rich and readily available sources of data and texts, whether stored in databases, newspapers, or in other scientific and business repositories. This has created the urgent need for novel computational theories and tools to analyse and extract valuable hidden insights from this explosive growth of digital data. Data mining, which extracts knowledge from structured datasets, and text mining which analyses unstructured documents, are subfields of knowledge discovery.

The stock market is a significant sector of a country's economy and represents a crucial role in the growth of their commerce and industry. Hence, discovering efficient ways to analyse and visualise stock market data is considered a significant issue in modern finance. Consequently, countries around the world depend on stock markets for their economic growth. Unfortunately, stock market crashes are unavoidable and are, by nature, preceded by speculative economical bubbles. The increasing importance of stock markets and their direct influence on the economy were the main reasons for deciding to study and analyse stock market crashes as the application domain for this research.

The need to determine early warning indicators for both banking and stock market crises has been the focus of study by many economists and politicians. Whilst most projects researching these critical indicators applied data mining using structured historical market prices to uncover hidden knowledge, very few attempted to adopt a text mining approach. Patel *et al.* (2015) explained that stock markets behave randomly; consequently, the application of data mining to the analysis of stock market data are

constrained to make assessments within the scope of existing information, and thus they are not able to model any random behaviour of stock market or provide causes behind events. One area of limited success in stock market prediction comes from textual data, which is a rich source of information and analysing it may provide better understanding of random behaviours of the market. Textual data limits the success in the investigation of stock markets because natural language is ambiguous, subtle and very rich. Given the huge amounts of free news and financial data, it is important to study the rich information embedded in this data, known as “alpha”. This research is an attempt at addressing this issue and discovers the critical indicators from unstructured yet valuable source of information.

Text mining is the focus of this research aimed at demonstrating its potential and valuable contribution to stock market crashes analysis, which is an important event of today’s global economy. Text mining involves the pre-processing of document collections, the extraction and representation of relevant features, the application of appropriate data mining techniques to analyse these intermediate representations through the application of supervised/unsupervised algorithms on these representations to discover new knowledge. Random forest classifier (supervised learning) has a number of strengths, which makes it worthwhile to further investigate and apply to analysis stock markets articles. RF can be a good predictor of stock markets because it uses ensemble strategies and random sampling. It is also less responsive to outlier data in training data and the bootstrapping and ensemble scheme help RF overcoming over fitting. These features have motivated this research to adopt RF and investigate its effectiveness in identifying critical indicators and evaluating their semantic contribution to the stock market movements. The goal of clustering (unsupervised learning), which is deployed in the implementation of SMRF-TM approach through the application of expectation maximisation is to distribute a set of data records into groups having high similarity. The expectation maximisation algorithm adds the objects to predefined clusters by calculating

their membership probabilities and follows this by updating the mixture model parameter in the maximisation step until the stopping criterion is reached.

The application area for this research is to text mine the 2009 Dubai stock market debt standstill. Some crashes, such as the 1929 Wall Street crash, can often be difficult to collect sufficient textual data (financial news) suitable for deep analysis. Stock market movements can be specific to particular economies and political environments such as the 1973-1974 United Kingdom stock market crash, the 1998 Russian financial crisis and the Chinese stock bubble of 2007. In 2009, a number of factors contributed to the United Arab Emirates crisis; these include the global recession, the bursting of the Dubai property bubble, and the post Lehman shutdown of international capital markets hit simultaneously. Dubai witnessed a significant slowdown in growth and strains in its banking system as a result of the global financial crisis, the decline in oil prices, and in particular the bursting of its property bubble.

6.2 Research Contributions

This research claims a number of contributions, which are described below.

1. The application of text mining to analyse rich information embedded in financial news related to stock markets to elicit critical indicators is an important contribution as most previous research projects focused on data mining analysing numerical data.

2. The application of text mining combined with Random Forest and expectation maximisation algorithms offers a novel approach to study these critical indicators, which can not only contribute to the prediction of stock market abnormal movements but also can enhance the performance of current trading systems' strategies.

a. The study demonstrates that Random Forest has outperformed the other classifiers and has achieved the best accuracy in classifying the financial news articles and the features extracted from the corpus.

b. The classification results of phase one of the experimental works show that the tree classifiers (ADTree and RF) outperformed the other types of tested classifiers, which are meta classifiers (e.g. Bagging), rules classifier (e.g. Decision Table) and Bayes classifier (e.g. Bayes Net). This indicates that the tree classifiers were more capable than the other types of classifiers in retrieving hidden information and the important relations between the extracted features, which supported better performance. These results support the first hypothesis of this research, which states that the application of the RF to the domain of stock market textual data can elicit the crucial relationships between the extracted features leading to the enhancement of the classification performance, and, therefore, it can be an effective predictive measure for the stock market movements.

c. In phase two of the experimental works, the 544 articles in the dataset corpus were used to check the effect of expanding the dataset on the performance of SMRF-TM system. The results showed that the tree classifiers (RF and Random Tree) perform much better than the Bayes classifier (Bayes Net), rules classifier (Decision Table) and the meta classifiers (Bagging and Rotation Forest) when applied on large datasets.

d. The experimentation results of phase two experiment two specifically, showed that bigrams based features/tokens enhanced the classification performance of SMRF-TM compared to the outcome achieved while using unigrams based features/tokens in phase one and phase two experiment one.

3. This research has developed a semi-supervised natural language processing driven approach, which is called Stock Market Random Forest-Text Mining system (SMRF-TM) to mine and extend the current classification of these critical indicators and their corresponding articles into new semantic classes.

    a. The random forest algorithm is applied to extend the classification of the extracted features and their articles from three to five classes: critical down, down, neutral, up and critical up.

    b. The expectation maximisation algorithm is applied to classify them further into three semantic classes: economic, social and political, thus extending current approaches from three to eight classes.

    c. This supported the second hypothesis of this research, which is that by extending the classification classes to five classes in SMRF-TM, the classification performances of both the extracted features as well as the news articles are enhanced. Hence, this extension also contributes to the novel approach of SMRF-TM.

    d. The application of expectation maximisation clustering technique to cluster the extracted features and the financial news articles according to their semantic meanings has also help understanding of the causes behind random forest classification for the features and the news articles.

6.3 Challenges and Limitations

This research faced some challenges, which are resolved and it also highlighted some limitations, which should be considered in future work. Feldman and Sanger (2007) and Yu *et al.* (2005) stated that text mining requires a lot of human input because of the need for continuous feedback from the domain experts to evaluate the results since the results may require further refining, as the final solutions may be sometimes uncertain, vague

and imprecise. Consequently, at the early stages of this research we needed the opinions of domain experts about the data source to be used for retrieving the required dataset for the purpose of the analysis of Dubai Debt stand still 2009 in order to make sure that it is a reliable source. Also, we needed the experts' opinions during the period of designing the SMRF-TM approach to ensure that we have encountered most of the significant tasks, which can support in the analysis of Dubai's stock market while implementing the SMRF-TM approach. Last but not least, the qualitative validation of the extracted features against the domain experts' opinions was very important in this research to check wither the extracted features and their relationships are appropriate for the analysis of Dubai's stock market domain or not.

Hence, the need to search and find the appropriate domain experts, who had the desired domain knowledge and availability was the first challenge faced at the early stages of this research. This challenge required some time and effort to find the appropriate experts for the study domain of this research. As discussed in chapter 5, this required a thorough investigation/search of the different fields related to the study domain of our research, which are the stock market, academic and banking fields and required interviews/meetings to identify the most appropriate experts, who had a good domain knowledge and experience as well as an accessible geographical location. This task successfully identified one academic expert in finance, one banking expert and two Stock Market experts.

Regarding the quantitative validation process, we faced a challenge, which was related to the software used to apply the random forest classifier. WEKA, the software used to apply the RF classifier, has a limited memory allocation and this caused a problem during experimentation, particularly in the experiments of phase two, which applied the complete dataset. The use of a big dataset (i.e. 544 news articles) caused the WEKA to run out of memory while applying some of the tested classifiers in phase two of the experimental

works such as ADTree, j48graft and Rotation Forest. In order to overcome this challenge a bigger memory size had to be allocated manually before each run.

Regarding the qualitative validation process, we faced another challenge, related to the inability of the experts to supply us with a reference list of critical indicators (features) related to the stock market movements. Consequently, the qualitative validation process had to be an interactive and lengthy process in order to make sure that the extracted features are critical indicators, which may have profound influence on the stock market movements.

One limitation was identified related to the data set. As discussed in chapter one, finding sufficient text data was a significant challenge, particularly in relation to older stock market crashes. Moreover, it was appropriate to choose a stock market where the domain experts had relevant expertise. For these reasons, the 2009 Dubai stock market debt standstill was chosen as the specific application domain for this research as data could be obtained from the Financial Times articles and the domain experts had expertise in the middle East stock market.

The reliance on Financial Times as the sole data source to train the SMRF-TM system is another limitation in this research, which should be addressed in future work. This limitation affected the results of this research in two ways. The first is that the random forest was only trained to retrieve the relationships between the extracted features and this therefore reflected the financial news reporting and presentation adopted by Financial Times. The second limitation was identified while applying the expectation maximisation clustering technique. Expectation maximisation was used to cluster the classified features and the news articles in one of three clusters: economic, social or political. The results found that most elements were clustered as economic and few were clustered as social or political because the dataset used was retrieved from a financial source only.

6.4 Future Work

The reasons behind financial crashes differ from one country to another and from one content to another. Hence, it is recommended for future work to apply the SMRF-TM system to study and analyse other financial crises, which occurred in different countries and to use different sources of financial news articles in order to refine the discovery of new critical indicators having different relationships. The employment of a variety of financial data sources to train the SMRF-TM system will support the use of random forest to extract the hidden information and relationships, which capture different interpretations from different data sources of the stock market news. However, this requires success in retrieving enough textual data suitable for analysis and finding appropriate financial experts with have the desired expertise in the specific market of study.

One positive outcome for the current implementation is the financial offer support from the managing director expert of a multinational brokering company, to extend SMRF-TM dataset to include the company's subscriptions to different financial news sources such as Bloomberg, Reuters, Wall Street Journal, and Economic Times. However, the causes of financial crashes are not always financial so relying only on financial data sources is not enough to provide the causes behind events. Consequently, in order to enhance the performance of the expectation maximisation clustering algorithm in clustering the classified features and the news articles according to their semantic meanings in one of the three clusters: economic, social or political, it is required not only to rely on financial data sources but also to include social news as well as political news.

References

ALADAG, C.H., YOLCU, U., EGRIOGLU, E. and BAS, E. (2014). Fuzzy lagged variable selection in fuzzy time series with genetic algorithms. *Applied Soft Computing*, *22*, pp.465-473.

ALI, J., KHAN, R., AHMAD, N. and MAQSOOD, I. (2012). Random Forests and Decision Trees. In *International Journal of Computer Science Issues,* Vol. 9, Issue 5, No 3.

ALI, M.M.Z. and THEODOULIDIS, B. (2014). Analyzing Stock Market Fraud Cases Using a Linguistics-Based Text Mining Approach. In *WaSABi-FEOSW@ ESWC*.

ANANIADOU, S., KELL, D.B. and TSUJII, J. (2006). Text Mining and Its Potential Applications in Systems Biology, *Trends in Biotechnology*, 24, 571-579.

ANASTASAKIS, L. and MORT, N. (2009). Exchange rate forecasting using a combined parametric and nonparametric self-organising modelling approach. *Expert Systems with Applications*, 36(10), pp.12001-12011.

APTE, C., DAMERAU, F. and WEISS, S.M. (1994). Automated Learning of Decision Rules for Text Categorization, *ACM Transactions on Information Systems*.

AREVALILLO, J. M. and NAVARRO, H. (2011). Uncovering Bivariate Interactions in High Dimensional Data Using Random Forests with Data Augmentation. *Fundamenta Informaticae*, *113*(2), 97-115.

ARLOT, S. and CELISSE, A. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys*, 4, pp.40-79.

BAEZA-YATES, R., BLANCO, R. and MALÚ CASTELLANOS, M. (2019). Web text mining. In Mitkov R. (Ed.) *Recent topics in NLP*. Oxford Handbook of Computational Linguistics, second edition. Oxford University Press (forthcoming publication).

BAHREPOUR, M., AKBARZADEH-T, M.R., YAGHOOBI, M. and NAGHIBI-S, M.B. (2011). An adaptive ordered fuzzy time series with application to FOREX. *Expert Systems with Applications*, 38(1), pp.475-485.

BAKER, L.D. and McCALLUM, A.K. (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*, pp. 96-103.

BARROW, D.K. and CRONE, S.F. (2016). Cross-validation aggregation for combining autoregressive neural network forecasts. *International Journal of Forecasting*, 32(4), pp.1120-1137.

BASALTO, N., BELLOTTI, R., DE CARLO, F., FACCHI, P. and PASCAZIO, S. (2005). Clustering stock market companies via chaotic map synchronization. *Physica A: Statistical Mechanics and its Applications*, *345*(1), 196-206.

BIAU, G., DEVROYE, L. and LUGOSI, G. (2008). Consistency of random forests and other averaging classifiers. *The Journal of Machine Learning Research*, *9*, 2015-2033.

BOLLEN, J., MAO, H. and ZENG, X. (2011). Twitter mood predicts the stock market. *Journal of computational science*, *2*(1), pp.1-8.

BOMFIM, A. (2000). Pre-announcement Effects, News, and Volatility: Monetary Policy and The Stock Market. *Technical report, Federal Reserve System*.

BOXELL, L. (2015). K-fold cross-validation and the gravity model of bilateral trade. *Atlantic Economic Journal*, 43(2), pp.289-300.

BRADLEY, P.S., FAYYAD, U.M. and REINA, C.A. (2000). Clustering very large databases using EM mixture models. In *Proceedings of 15th International Conference on Pattern Recognition.* Vol. 2, pp. 76-80. IEEE.

BREIMAN, L. (2001). Random forests. *Machine Learning*, 45(1): 5-32.

CARVALHOB, C., KLAGGEA, N. and MOENCHA, E. (2011). The Persistent Effects of a False News Shock. Journal of Empirical Finance, 18(4):597-615.

CHAKRABORTY, R. (2013). DOMAIN KEYWORD EXTRACTION TECHNIQUE: A New WEIGHTING METHOD. *Computer Science & Information Technology*, 109.

CHEEMA, A., VORA, A., JAIN, C., KATARIA, P., SHAH, R. and WAGH, S. (2008). Stock Forecasters.

CHEN, H., ZHAN, Y. and LI, Y. (2010). The application of decision tree in Chinese email classification. In *Machine Learning and Cybernetics (ICMLC), 2010 International Conference*, Vol. 1, pp. 305-308. IEEE.

CHUA, W. F. (1986). Radical Developments in Accounting Thought, *The Accounting Review*, LXI:4, 601-632.

CHURCH, K. and GALE, W. (1999). Inverse document frequency (IDF): A measure of deviations from poisson. In *Natural language processing using very large corpora* (pp. 283-295). Springer Netherlands.

CROFT, W. B. and HARPER, D. J. (1979). Using probabilistic models of document retrieval without relevance information. *Journal of documentation*, *35*(4), 285-295.

CULTER, D., POTERBA, J. and SUMMERS, L. (1991). Speculative Dynamics. *The Review of Economic Studies*, 529-546.

CUNNINGHAM, P. (2007). *Ensemble techniques*. Technical Report UCD-CSI-2007–5.

DAVIS, A., PIGER, J. and SEDOR, L. (2006). Beyond the Numbers: An Analysis of Optimistic and Pessimistic Language in Earnings Press Releases. *Technical report, Federal Reserve Bank*.

DE BONDT, W. F. M. and THALER, R. H. (1985). Does The Stock Market Over React? *Journal of Finance*, 40:793-805.

DEHMER, M., GRABNER, M. and VARMUZA, K. (2012). Information indices with high discriminative power for graphs. *PLoS One*, *7*(2), p.e31214.

DENG, Z., ZHU, X., CHENG, D., ZONG, M. and ZHANG, S. (2016). Efficient kNN classification algorithm for big data. *Neurocomputing*, *195*, pp.143-148.

DIETTERICH, T. G. (2002). Ensemble Learning. *The Handbook of Brain Theory and Neural Networks*.

DONATE, J.P., CORTEZ, P., SANCHEZ, G.G. and De MIGUEL, A.S. (2013). Time series forecasting using a weighted cross-validation evolutionary artificial neural network ensemble. *Neurocomputing*, *109*, pp.27-32.

DRURY, B. (2013). A Text Mining System for Evaluating the Stock Market's Response To News, *Doctoral Program in Computer science of the Universities of Minho, Aveiro and Porto*.

DUMAIS, S. T. (2004). Latent semantic analysis. *Annual review of information science and technology*, *38*(1), 188-230.

ESTRADA, F.J. and JEPSON, A.D. (2009). Benchmarking image segmentation algorithms. *International Journal of Computer Vision*, *85*(2), pp.167-181.

FAN, Y., SHEN, D. and DAVATZIKOS, C. (2005). Classification of structural images via high-dimensional image warping, robust feature extraction, and SVM. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2005*, pp.1-8.

FARMER, R.E. (2015). The stock market crash really did cause the great recession. *Oxford Bulletin of Economics and Statistics*, *77*(5), pp.617-633.

FELDMAN, R. and DAGAN, I. (1995). Knowledge Discovery in Textual Databases (KDT). In *KDD* (Vol. 95, pp. 112-117).

FELDMAN, R. and SANGER, J. (2007). *The Text Mining Handbook,* New York, Cambridgre University Press.

FENG, J., XIE, F., HU, X., LI, P., CAO, J. and WU, X. (2011). Keyword Extraction Based on Sequential Pattern Mining, *3rd International Conference on Internet Multimedia Computing and Service*, 34-38.

FORMAN, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of machine learning research*, *3*(Mar), pp.1289-1305.

FORMAN, G. (2007). Feature selection for text classification. In *Computational methods of feature selection*, Chapman and Hall/CRC Press. *1944355797*.

GABLE, G.G. (1994). Integrating case study and survey research methods: an example in information systems. *European journal of information systems*, *3*(2), pp.112-126.

GÁLVEZ, R.H. and GRAVANO, A. (2017). Assessing the usefulness of online message board mining in automatic stock prediction systems. *Journal of Computational Science*, *19*, pp.43-56.

GEISSER, S. (1975). The predictive sample reuse method with applications. *Journal of the American statistical Association*, *70*(350), pp.320-328.

GENG, R., BOSE, I. and CHEN, X. (2015). Prediction of financial distress: An empirical study of listed Chinese companies using data mining. *European Journal of Operational Research*, *241*(1), pp.236-247.

GHAFFARI, N., YOUSEFI, M. R., JOHNSON, C. D., IVANOV, I., and DOUGHERTY, E. R. (2013). Modeling the next generation sequencing sample processing pipeline for the purposes of classification. BMC bioinformatics, 14(1), 307.

GHAZALI, R., HUSSAIN, A.J. and LIATSIS, P. (2011). Dynamic Ridge Polynomial Neural Network: Forecasting the univariate non-stationary and stationary trading signals. *Expert Systems with Applications*, *38*(4), pp.3765-3776.

GHOSH, S., ROY, S. and BANDYOPADHYAY, S. K. (2012). A Tutorial Review on Text Mining Algorithms. In *International Journal of Advanced Research in Computer and Communication Engineering,* Vol. 1, Issue 4.

GLUCKSBERG, S. (2008). Sentiment Analysis: Emotion, Metaphor, Ontology and Terminology. In *Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, Proceedings of Language Resources and Evaluation (LREC)*, pp. 94-101.

GÓMEZ, M. M. Y., GELBUKH, A. and LÓPEZ, A. L. (2001). Mining the News: Trends, Associations, and Deviations, *Computación Sistemas,* 5**,** 14-24.

GONÇALVES, P., ARAÚJO, M., BENEVENUTO, F. and CHA, M. (2013). Comparing and combining sentiment analysis methods. In *Proceedings of the first ACM conference on Online social networks*, pp. 27-38. ACM.

GOODWIN, L.D. and LEECH, N.L. (2003). The Meaning of Validity in the New Standards for Educational and Psychological Testing: Implications for Measurements Courses. *Measurement and evaluation in Counseling and Development*, 36(3), pp.181-91.

GREIFF, W. R. (1998). A theory of term weighting based on exploratory data analysis. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 11-19). ACM.

GROCHOLEWSKI, S. (1997). CORPORA-speech database for Polish diphones. In *Proceedings of Eurospeech*.

GUO, Y., BENNAMOUN, M., SOHEL, F., LU, M., WAN, J. and KWOK, N.M. (2016). A comprehensive performance evaluation of 3D local feature descriptors. *International Journal of Computer Vision*, 116(1), pp.66-89.

GUPTA, V. and LEHAL, G. S. (2009). A Survey of Text Mining Techniques and Applications, *Journal of Emerging Technologies in Web Intelligence*, 1, 1, 60-76.

HAFEZ, P. A. (2009). Construction of Market Sentiment Indices Using News Sentiment. *Technical Report, Ravenpack*.

HAFEZ, P. A. (2010). How News Events Impact Market Sentiment. *Technical report, Ravenpack*.

HAJIZADEH, E., ARDAKANI, H. D. and SHAHRABI, J. (2010). Application of data mining techniques in stock markets: A survey. *Journal of Economics and International Finance*, *2*(7), 109-118.

HAKIM, A.A., ERWIN, A., ENG, K.I., GALINIUM, M. and MULIADY, W. (2014). Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach. In *Information Technology and Electrical Engineering (ICITEE), 6th International Conference on*, pp.1-4. IEEE.

HAN, B., LIU, Y., GINZINGER, S.W. and WISHART, D.S. (2011). SHIFTX2: significantly improved protein chemical shift prediction. *Journal of biomolecular NMR*, 50(1), p.43.

HAN, J. and KAMBER, M. (2006). *Data Mining: Concepts and Techniques,* San Francisco, Morgan Kaufmann Publishers is an imprint of Elsevier.

HEARST, M. (2003). What is text mining. *SIMS, UC Berkeley*.

HENRY, E. (2006). Market Reaction to Verbal Components of Earnings Press Releases. *Journal of Emerging Technologies in Accounting*, 3:1-19.

HILLENMEYER, M. E., ERICSON, E., DAVIS, R. W., NISLOW, C., KOLLER, D. and GIAEVER, G. (2010). Method Systematic analysis of genome-wide fitness data in yeast reveals novel gene function and drug action. *Genome biology*, *11*(3).

HIRSCHHEIM, R. (1985). Information systems epistemology: An historical perspective. *Research methods in information systems*, pp.13-35.

HONG, H. and STEIN, J.C. (1999). A unified theory of underreaction, momentum trading, and overreaction in asset markets. *The Journal of finance*, *54*(6), pp.2143-2184.

HORNING, N. (2013). Introduction to decision trees and random forests. *American Museum of Natural History's*.

HUANG, C., TIAN, Y., ZHOU, Z., LING, C. X. and HUANG, T. (2006). Key phrase Extraction Using Semantic Networks Structure Analysis, *6th International Conference on Data Mining,* 275-284.

JABEEN, A., AHMAD, N. and RAZA, K. (2018). Machine Learning-Based State-of-the-Art Methods for the Classification of RNA-Seq Data. In *Classification in BioApps* (pp. 133-172). Springer, Cham.

JANECEK, A. (2009). *Efficient feature reduction and classification methods* (Doctoral dissertation, uniwien).

JIANG, P. and CHEN, J. (2016). Displacement prediction of landslide based on generalized regression neural networks with K-fold cross-validation. *Neurocomputing*, *198*, pp.40-47.

JOHNSON, D.E., OLES, F.J., ZHANG, T. and GOETZ, T. (2002). A decision-tree-based symbolic rule induction system for text categorization. *IBM Systems Journal*, *41*(3), pp.428-437.

JUNG, Y.G., KANG, M.S. and HEO, J. (2014). Clustering performance comparison using K-means and expectation maximization algorithms. *Biotechnology & Biotechnological Equipment*, *28*(sup1), pp.S44-S48.

KAMRUZZAMAN, S.M., HAIDER, F. and HASAN, A.R. (2010). Text classification using data mining. *arXiv preprint arXiv:1009.4987*.

KANNAN, K.S., SEKAR, P.S., SATHIK, M.M. and ARUMUGAM, P. (2010). Financial Stock Market Forecast using Data Mining Techniques, *International MultiConference of Engineers and computer scientists.* Hong Kong,1.

KANNAN, K.S. and GURUSAMY, V. (2014). Preprocessing Techniques for Text Mining.

KAPLAN, B. and MAXWELL, J.A. (1994). Evaluating health care information systems: Methods and applications. *Qualitative Research Methods for Evaluating Computer Information Systems.* JG Anderson, CE Ayden and SJ Jay. Thousand Oaks, Sage.

KAYA, M.I.Y. and KARSLIGIL, M.E. (2010). Stock Price Prediction Using Financial News Articles, *2nd IEEE International Conference on Information and Financial Engineering,* 478 - 482.

KHEDR, A.E., SALAMA, S.E. and YASEEN, N. (2017). Predicting Stock Market Behavior using Data Mining Technique and News Sentiment Analysis. *International Journal of Intelligent Systems and Applications (IJISA)*, 9(7), 22-30.

KIM, S.B., HAN, K.S., RIM, H.C. and MYAENG, S.H. (2006). Some effective techniques for naive bayes text classification. *IEEE transactions on knowledge and data engineering*, *18*(11), pp.1457-1466.

KIM, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:*1408.5882.

KIRKOS, E., SPATHIS, C. and MANOLOPOULOS, Y. (2007). Data mining techniques for the detection of fraudulent financial statements. *Expert systems with applications*, *32*(4), pp.995-1003.

KLOPTCHENKO, A., EKLUND, T., BACK, B., KARLSSON, J., VANHARANTA, H. and VISA, A. (2002). Combining Data and Text Mining Techniques for Analyzing Financial Reports, *8th Americas Conference on Information Systems,* 20-28.

KOHAVI, R. and PROVOST, F. (1998). Confusion matrix. *Machine learning*, *30*(2-3), pp.271-274.

KORDE, V. and MAHENDER, C.N. (2012). Text classification and classifiers: A survey. *International Journal of Artificial Intelligence & Applications*, *3*(2), p.85.

KOTSIANTIS, S.B., ZAHARAKIS, I. and PINTELAS, P. (2007). Supervised machine learning: A review of classification techniques. Informatica 31 (3) 249-268.

KUMAR, N. and KHATRI, S. (2017). Implementing WEKA for medical data classification and early disease prediction. In *Computational Intelligence & Communication Technology (CICT), 2017 3rd International Conference on* (pp. 1-6). IEEE.

KUMAR, N. and KHATRI, S. (2017). Significance of Data Mining in Disease Classification and Prediction for Mining Clinical Data: A Review. *International Journal of Advanced Research in Computer Science*, *8*(5).

KUNTRARUK, J. and POTTENGER, W. M. (2001). Massively Parallel Distributed Feature Extraction in Textual Data Mining Using HDDI™, *10th IEEE International Symposium on High Performance Distributed Computing,* 363-370.

LARSON, S.C. (1931). The shrinkage of the coefficient of multiple correlation. *Journal of Educational Psychology*, *22*(1), p.45.

LAZAREVIC-McMANUS, N., RENNO, J.R., MAKRIS, D. and JONES, G.A. (2008). An object-based comparative methodology for motion detection based on the F-Measure. *Computer Vision and Image Understanding*, *111*(1), pp.74-85.

LEBART, L. (2004). Validation Techniques in Text Mining (with Application to the Processing of Open-ended Questions). In *Text Mining and Its Applications* (pp. 169-178). Springer, Berlin, Heidelberg.

LEE, H., SURDEANU, M., MACCARTNEY, B. and JURAFSKY, D. (2014). On the Importance of Text Analysis for Stock Price Prediction. In *LREC* (pp. 1170-1175).

LEWIS, D.D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. In *European conference on machine learning* (pp. 4-15). Springer, Berlin, Heidelberg.

L'HUILLIER G., HEVIA, A., WEBER, R. and R'IOS, S. (2010). Latent Semantic Analysis and Keyword Extraction for Phising Classification, *IEEE International Conference on Intelligence and Security Informatics,* 129-131.

LIANGTU, S. and XIAOMING, Z. (2007). Web Text Feature Extraction with Particle Swarm Optimization, *International Journal of Computer Science and Network Security*, 7, 6, 132-136.

MABU, S., HIRASAWA, K., OBAYASHI, M. and KUREMOTO, T. (2013). Enhanced decision making mechanism of rule-based genetic network programming for

creating stock trading signals. *Expert Systems with Applications*, *40*(16), pp.6311-6320.

MAHAJAN, A., DEY, L. and HAQUE, S. M. (2008). Mining Financial News for Major Events and Their Impacts on the Market, In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT.* Sydney, Volume (1), (pp. 423-426). IEEE Computer Society.

MAHGOUB, H., RÖSNER, D., ISMAIL, N. and TORKEY, F. (2008). A text mining technique using association rules extraction. *International journal of computational intelligence*, *4*(1), pp.21-28.

MARATEA, A., PETROSINO, A. and MANZO, M. (2014). Adjusted F-measure and kernel scaling for imbalanced data learning. *Information Sciences*, *257*, pp.331-341.

MARKUS, M.L. (1994). Electronic mail as the medium of managerial choice. *Organization science*, *5*(4), pp.502-527.

McCALLUM, A. and NIGAM, K. (1998). A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization* (Vol. 752, pp. 41-48).

McMANUS, J. (1988). An Economic Theory of News Selection. In *Annual Meeting for Education in Journalism and Mass Communication*.

MEHTA, M., AGRAWAL, R. and RISSANEN, J. (1996). SLIQ: A fast scalable classifier for data mining. *Advances in Database Technology—EDBT'96*, pp.18-32.

MINER, G. (2012). Practical text mining and statistical analysis for non-structured text data applications. *Academic Press*.

MING, F., WONG, F., LIU, Z. and CHIANG, M. (2014). Stock market prediction from WSJ: text mining via sparse matrix factorization. In *International Conference on Data Mining (ICDM),* (pp. 430-439).

MINGERS, J. (2001). Combining IS research methods: towards a pluralist methodology. *Information systems research*, *12*(3), pp.240-259.

MISHKIN, F.S. and WHITE, E.N. (2002). US stock market crashes and their aftermath: implications for monetary policy (No. w8992). *National bureau of economic research*.

MITTERMAYER, M. A. and KNOLMAYER, G. (2006). Text Mining Systems for Predicting Market Response to News: A Survey, *Arbeitsbericht Nr. 184 des Institut für Wirtschaftsinformatik der Universitat Bern***,** WP-184.

MITRA, G. and MITRA, L. (2011). The Handbook of News Analytics in Finance, chapter "How news events impact market sentiment", pages 129-145. Wiley Finance.

MORENO-TORRES, J.G., SÁEZ, J.A. and HERRERA, F. (2012). Study on the impact of partition-induced dataset shift on k-fold cross-validation. *IEEE Transactions on Neural Networks and Learning Systems*, *23*(8), pp.1304-1312.

MOSTELLER, F. and TUKEY, J.W. (1968). Data Analysis, Including Statistics In: G. Lindzey and E. Aronson, editors, Handbook of Social Psychology, Vol. 2.

MYERS, M. D. (1997). Qualitative research in information systems. *Management Information Systems Quarterly*, *21*, 241-242.

MYUNG, J., YANG, J. Y. and LEE, S. G. (2009). PicAChoo: A Tool for Customizable Feature Extraction Utilizing Characteristics of Textual Data, *3rd International Conference on Ubiquitous Information Management and Communication*, 650-655.

NAKHAEIZADEH, G., STEURER, E. and BARTLMAE, K. (2002). Banking and finance. In *Handbook of data mining and knowledge discovery* (pp. 771-780). Oxford University Press, Inc..

NASSIRTOUSSI, A.K., AGHABOZORGI, S., WAH, T.Y. and NGO, D.C.L. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications*, 41(16), 7653-7670.

NASSIRTOUSSI, A.K., AGHABOZORGI, S., WAH, T.Y. and NGO, D.C.L. (2015). Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment. *Expert Systems with Applications*, 42(1), pp.306-324.

NASUKAWA, T. and NAGANO, T. (2001). Text Analysis and Knowledge Mining System, *IBM Systems Journal,* 40,4, 967-984.

NARDO, M., PETRACCO-GIUDICI, M. and NALTSIDIS, M. (2016). Walking down Wall Street with a tablet: A survey of stock market predictions using the Web. *Journal of Economic Surveys*, 30(2), 356–369.

NIEDERHOFFER, V. (1971). The Analysis of World Events and Stock Prices. *Journal Of Business,* 44(2):193-219.

NIKFARJAM, A., EMADZADEH, E. and MUTHAIYAH, S. (2010). Text Mining Approaches for Stock Market Prediction, *2nd International Conference on Computer and Automation Engineering (ICCAE)*, 4, 256-260.

ORLIKOWSKI, W.J. and BAROUDI, J.J. (1991). Studying information technology in organizations: Research approaches and assumptions. *Information systems research*, *2*(1), pp.1-28.

PATEL, J., SHAH, S., THAKKAR, P. and KOTECHA, K. (2015). Predicting stock market index using fusion of machine learning techniques. *Expert Systems with Applications*, 42(4), 2162-2172.

PATEL, N. (2017). An Optimized Classifier Frame Work based on Rough Set and Random Tree. *International Journal of Computer Applications*, *160*(9).

PERCHA, B., GARTEN, Y. and ALTMAN, R. B. (2012). Discovery and explanation of drug-drug interactions via text mining. In *Pac Symp Biocomput* (Vol. 410, p. 421).

PONCIANO, R., PAIS, S. and CASAL, J. (2015). Using accuracy analysis to find the best classifier for Intelligent Personal Assistants. *Procedia Computer Science*, *52*, pp.310-317.

PÖNKÄ, H. (2017). Predicting the direction of US stock markets using industry returns. *Empirical Economics*, *52*(4), pp.1451-1480.

PREMANODE, B., VONPRASERT, J. and TOUMAZOU, C. (2013). Prediction of exchange rates using averaging intrinsic mode function and multiclass support vector regression. *Artificial Intelligence Research*, *2*(2), p.47.

PRINZIE, A. and POEL, D. V. D. (2008). Random forests for multiclass classification: Random multinomial logit. *Expert systems with Applications*, *34*(3), 1721-1732.

QI, Y. (2011). Random Forest for Bioinformatics.

www.cs.cmu.edu/~qyj/papersA08/11-rfbook.pdf, (accessed on September 2017)

RIJSBERGEN, C.J. (1979). Information retrieaval: Introduction. online book http://www. dcs. gla. ac. uk/Keith/Chapter. 1.

RISH, I. (2001). An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence* (Vol. 3, No. 22, pp. 41-46). IBM.

ROBBANI, M. and ANANTHARAMAN, S. (2004). An Econometric Analysis of Stock Market Reaction to Political Events in Emerging Markets. *In Proceedings of Second Annual ABIT Conference*.

RODRIGUEZ, J.D., PEREZ, A. and LOZANO, J.A. (2010). Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE transactions on pattern analysis and machine intelligence*, *32*(3), pp.569-575.

ROMO, J. M. and ARAUJO, L. (2013). Detecting Malicious Tweets in Trending Topics Using A Statistical Analysis of Language. *Expert Systems with Applications 40, 2992–3000.*

SALEHI, M., MOUSAVI, S. M. and BOLANDRAFTAR, P. M. (2016). Predicting corporate financial distress using data mining techniques: An application in

Tehran Stock Exchange. *International Journal of Law and Management*, *58*(2), pp.216-230.

SAMMUT, C. and WEBB, G.I. (2011). *Encyclopedia of machine learning*. Springer Science & Business Media.

SCHUMAKER, R. P. and CHEN, H. (2009). Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFinText System, *ACM Trans. Inf. Syst.*, 27, 2, 12:1--12:19.

SCHUMAKER, R. P., ZHANG, Y., HUANG, C. N. and CHEN, H. (2012). Evaluating sentiment in financial news articles. *Decision Support Systems*, *53*(3), 458-464.

SCHUSTER, T. (2003). News Events and Price Movements. Price Effects of Economic and Noneconomic Publications in The News Media. Technical report, EconWPA.

SERMPINIS, G., THEOFILATOS, K., KARATHANASOPOULOS, A., GEORGOPOULOS, E.F. and DUNIS, C. (2013). Forecasting foreign exchange rates with adaptive neural networks using radial-basis functions and particle swarm optimization. *European Journal of Operational Research*, *225*(3), pp.528-540.

SHEN, Y. and JIANG, J. (2003). Improving the performance of Naive Bayes for text classification. *CS224N Spring*.

SORTO, M., AASHEIM, C. and WIMMER, H. (2017). Feeling The Stock Market: A Study in the Prediction of Financial Markets Based on News Sentiment. In: *Proceedings of the Southern Association for Information Systems Conference*. St. Simons Island, GA, USA.

SPENCER, R. (2009). Dubai's Financial Crisis. www.telegraph.co.uk, (accessed on July 2017)

STEWART, B. M. and ZHUKOV, Y. M. (2009). Use of Force and Civil–Military Relations in Russia: An Automated Content Analysis. *Small Wars & Insurgencies* Vol. 20, No. 2, 319–343.

STONE, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society. Series B (Methodological)*, pp.111-147.

STREINER, D.L. (2003). Starting at the beginning: an introduction to coefficient alpha and internal consistency. *Journal of personality assessment*, *80*(1), pp.99-103.

STREINER, D.L. and NORMAN, G.R. (2006). “Precision” and “accuracy”: two terms that are neither. *Journal of clinical epidemiology*, *59*(4), pp.327-330.

SUN, A., LACHANSKI, M. and FABOZZI, F.J. (2016). Trade the tweet: Social media text mining and sparse matrix factorization for stock market prediction. *International Review of Financial Analysis*, 48, 272-281.

SUN, J. and LI, H. (2008). Data mining method for listed companies’ financial distress prediction. *Knowledge-Based Systems*, *21*(1), 1-5.

SVETNIK, V., LIAW, A., TONG, C., CULBERSON, J. C., SHERIDAN, R. P. and FEUSTON, B. P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of chemical information and computer sciences*, *43*(6), 1947-1958.

SVETNIK, V., LIAW, A., TONG, C. and WANG, T. (2004). Application of Breiman’s random forest to modeling structure-activity relationships of pharmaceutical molecules. In *Multiple Classifier Systems* (pp. 334-343). Springer Berlin Heidelberg.

TAN, A.H. (1999). Text mining: The state of the art and the challenges. In *Proceedings of the PAKDD Workshop on Knowledge Disocovery from Advanced Databases* (Vol. 8, pp. 65-70). sn.

TAŞCI, S. and GÜNGÖR, T. (2008). An evaluation of existing and new feature selection metrics in text categorization. In *Computer and Information Sciences, 2008. ISCIS'08. 23rd International Symposium on* (pp. 1-6). IEEE.

TAŞCI, Ş. and GÜNGÖR, T. (2013). Comparison of text feature selection policies and using an adaptive framework. *Expert Systems with Applications*, *40*(12), pp.4871-4886.

TETLOCK, P., TSECHANSKY, M. S. and MACSKASSY, S. (2008). More Than Words: Quantifying Language to Measure Firms' fundamentals. *Journal of Finance*, 63:1437-1467.

TRIBA, M.N., Le MOYEC, L., AMATHIEU, R., GOOSSENS, C., BOUCHEMAL, N., NAHON, P., RUTLEDGE, D.N. and SAVARIN, P. (2015). PLS/OPLS models in metabolomics: the impact of permutation of dataset rows on the K-fold cross-validation quality parameters. *Molecular BioSystems*, *11*(1), pp.13-19.

UPPAL, K. and LEE, E.K. (2017). SEACOIN2. 0: an interactive mining and visualization tool for information retrieval, summarization, and knowledge discovery. *bioRxiv*, p.206193.

VANSTONE, B. and FINNIE, G. (2010). Enhancing stockmarket trading performance with ANNs. *Expert Systems with Applications*, *37*(9), pp.6602-6610.

VIDHYA, K.A. and AGHILA, G. (2010). Hybrid text mining model for document classification. In *The 2nd International Conference on Computer and Automation Engineering (ICCAE),* (Vol. 1, pp. 210-214). IEEE.

VIJAYARANI, S., ILAMATHI, M.J. and NITHYA, M. (2015). Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, *5*(1), pp.7-16.

WANG, S., XU, K., LIU, L., FANG, B., LIAO, S. and WANG, H. (2011). An Ontology Based Framework for Mining Dependence Relationships Between News and Financial Instruments, *Expert Systems with Applications,* 38**,** 12044-12050.

WEI, C. P. and DONG, Y. X. (2001). A mining-based category evolution approach to managing online document categories. In *System Sciences, 2001. Proceedings of the 34th Annual Hawaii International Conference on* (pp. 10-pp). IEEE.

WEISS, S.M., INDURKHYA, N. and ZHANG, T. (2010). *Fundamentals of predictive text mining* (Vol. 41). London: Springer.

WITTEN, I.H., FRANK, E., HALL, M.A. and PAL, C.J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

WONG T. L. and LAM, W. (2009). An Unsupervised Method for Joint Information Extraction and Feature Mining Across Different Web Sites, *Data and Knowledge Engineering*, 68, 1, 107-125.

WONG, T.T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, *48*(9), pp.2839-2846.

WONG, T.T. (2017). Parametric methods for comparing the performance of two classification algorithms evaluated by k-fold cross validation on multiple data sets. *Pattern Recognition*, *65*, pp.97-107.

WU, C.H. (2009). Behavior-based spam detection using a hybrid method of rule-based techniques and neural networks. *Expert Systems with Applications*, *36*(3), pp.4321-4330.

WU, M. C., LIN, S. Y. and LIN, C. H. (2006). An effective application of decision tree to stock trading. *Expert Systems with Applications*, *31*(2), 270-274.

WU, X., ZHU, X., WU, G.Q. and DING, W., 2014. Data mining with big data. *IEEE transactions on knowledge and data engineering*, *26*(1), pp.97-107.

WUTHRICH, B., PERMUNETILLEKE, D., LEUNG, S., CHO, V. and LAM, W. (1998). Daily Prediction of Major Stock Indices from Textual WWW Data, *4th International Conference on Knowledge Discovery and Data Mining, 364-368*.

XU, R. and WUNSCH, D. (2005). Survey of clustering algorithms. *IEEE Transactions on neural networks*, *16*(3), pp.645-678.

YU, D.J., HU, J., HUANG, Y., SHEN, H.B., QI, Y., TANG, Z.M. and YANG, J.Y. (2013). TargetATPsite: a template-free method for ATP-binding sites prediction

with residue evolution image sparse representation and classifier ensemble. *Journal of computational chemistry*, *34*(11), pp.974-985.

YU, L., WANG, S. and LAI, K. K. (2005). A rough-set-refined text mining approach for crude oil market tendency forecasting. *International Journal of Knowledge and Systems Sciences*, *2*(1), 33-46.

ZHAO, Y., CHEN, F., ZHAI, R., LIN, X., WANG, Z., SU, L. and CHRISTIANI, D. C. (2012). Correction for population stratification in random forest analysis. *International journal of epidemiology*, dys183.

ZIÓŁKO, B. (2015). Fuzzy precision and recall measures for audio signals segmentation. *Fuzzy Sets and Systems*, *279*, pp.101-111.

ZIÓŁKO, B., MANANDHAR, S. and WILSON, R.C. (2007). Fuzzy recall and precision for speech segmentation evaluation. In *Proceedings of 3rd Language and Technology Conference, Poznań*.

Appendix A

reduc	low	rate	grow	decad	current	loss
bond	market	equiti	term	risk	properti	look
real	calper	inflat	bank	know	privat	soon
fund	manag	need	earn	public	gradual	practic
vote	global	price	intern	dubai	industri	compani
strong	achiev	saudi	fall	jame	remain	sharehold
princ	demand	india	start	meet	foreign	parti
hedg	consum	china	suppli	think	expect	investor
asset	spend	job	new	websit	growth	consumpt
long	import	polici	cycl	project	countri	budget
high	recess	gener	larg	continu	monetari	centuri
group	experi	appreci	sector	deliveri	exampl	inventori
drug	war	particl	emerg	quickli	structur	discretionari
result	recent	cours	run	higher	perform	economi
gulf	hike	event	scale	debt	financi	unemploy
presid	propos	financ	region	labour	consolid	economist
reset	borrow	clear	pain	candid	investig	parliament
cut	boom	suprem	deflat	surpris	credibl	downgrad
battl	tighten	leverag	bubbl	exagger	lansdown	nakheel
crisi	shortag	defens	mayor	immigr	destroi	bridgepoint

Appendix A.1 Phase 1-Examples of features classified to class down

properti	firm	develop	plan	market	growth	economi
approach	fundament	profit	skill	increas	manag	residenti
continu	dealership	rate	bond	fund	polici	stronger
present	structur	favor	posit	privat	asset	product
practic	support	govern	hold	corpor	actual	investor
flexibl	contagion	equiti	trade	partner	tool	exhibit
industri	restructur	innov	forum	moder	secur	compani
media	economist	elect	audit	law	found	quicken
creat	governor	convent	arbitr	counti	adopt	interview
inventori	compliant	banker	institut	deriv	lawyer	legal
vessel	lender	ship	stock	termin	regul	bondhold
collect	museum	contain	prime	justic	advic	sharehold
interlaw	architect	brand	team	design	volum	modernist
project	guarante	automat	pearl	citadel	cultur	freshfield
founder	transgend	activist	unifi	middl	build	partnership
brilliant	collabor	logist	new	price	capit	sponsorship
privatis	collector	auction	intern	level	sale	contractor
immigr	brokerag	interior	matur	peopl	come	financ
region	comment	opposit	mix	busi	short	tighten
global	challeng	sector	liquid	inflat	need	current
public	exposur	countri	retir	data	work	differenti

Appendix A.2 Phase 1-Examples of features classified to class up

abu	dhabi	africa	aldar	asset	bank	compani
bond	busi	capit	centr	come	cultur	continu
data	dubai	econom	arab	elect	europ	current
financ	firm	fund	global	govern	group	happen
intern	gulf	equiti	know	Level	like	industri
list	make	manag	market	nation	need	project
new	oil	pearl	peopl	polici	plan	inventori
public	price	moodi	privat	immigr	flexibl	sector
level	like	rate	region	rent	sale	spend
state	stock	sukuk	term	trade	regul	tighten
team	time	thing	think	polit	prime	presid

Appendix A.3 Phase 1-Examples of features classified to class down and to class up

Appendix B

exchang	manag	trade	stock	technic	market	share
foreign	presid	develop	financi	bond	matur	bank
need	debt	conflict	price	intern	defianc	recess
regim	govern	asset	region	leader	econom	institut
sector	compani	financ	term	product	nakheel	dealer
sale	rate	wage	inflat	afford	candid	interior
elect	fund	report	investor	emerg	problem	grow
instanc	actual	index	number	construct	restructur	regul
purpos	budget	summit	discuss	travel	compliant	vote
industri	shortag	popul	job	state	properti	suppli
deficit	loan	credit	insur	dividend	manufactur	pursu
recruit	banker	destroi	convent	brought	reformist	left
moodi	liber	contain	employ	intellectu	assassin	trick
burn	consum	inflamm	basic	patent	downgrad	bail
death	crook	auction	buyer	risk	pendragon	illeg
collect	unknown	incorpor	captiv	collater	campaign	privat

Appendix B.1 Phase 2 Experiment 1-Examples of unigram features classified to class down

bond	investor	market	compani	share	provid	price
trade	financi	new	group	fund	bank	oil
privat	invest	secur	develop	list	adopt	defens
busi	founder	expect	review	manag	partner	growth
row	asset	retail	work	earn	properti	data
estat	local	client	lender	regul	approach	bring
corpor	govern	tadawul	clear	advic	lawyer	law
post	support	economi	construct	project	transport	land
feder	lend	schedul	consolid	council	leverag	lord
engag	advis	formula	expatri	round	profession	team
media	involv	reuter	sector	healthcar	sponsorship	engin
legal	propos	creditor	restructur	investig	sharehold	found
sukuk	afford	repaid	skill	collabor	partnership	repres
counti	activist	transfer	maker	worker	sectarian	pearl
invad	profess	consum	advertis	interlaw	transgend	centr

Appendix B.2 Phase 2 Experiment 1-Examples of unigram features classified to class up

time	south	nation	world	year	rate	cent
bond	privat	offic	need	term	like	expect
look	global	bank	learn	past	high	dubai
metal	saudi	europ	list	paper	gulf	histor
region	guess	agre	itali	hous	sub	busi
data	prime	manag	bullion	india	china	east
import	cultur	car	estim	chines	think	issu
central	london	state	open	mine	research	sale
make	work	russia	effect	websit	dollar	job
new	impact	tonn	opposit	group	number	secur
vehicl	result	parent	western	iranian	european	iran
airbu	airlin	africa	african	arab	tunisia	uae
resourc	bric	brazil	franc	tehran	modernist	fiscal
row	firm	british	languag	mean	washington	alloc
immigr	abu	dhabi	sharia	nasdaq	instrument	present

Appendix B.3 Phase 2 Experiment 1-Examples of unigram features classified to class neutral

dicreas	frontier	small	respons	crisi	problem	drawn
weak	persist	hedg	squeez	quit	fundament	crash
commod	larg	boom	sizeabl	rapid	debentur	posit
inflat	limit	debt	trap	risk	diversifica	notnorm
late	strategi	exagger	reduc	hit	withdraw	unemploy
deflat	bull	weaker	impact	far	negative	consequ
volatili	gradual	prone	spare	low	underwai	difficult
burst	lose	shortag	leav	littl	underli	tighten

Appendix B.4 Phase 2 Experiment 1-Examples of unigram features classified to class critical down

securiti	trade	mission	fund	econom	launch	recoveri
support	develop	financi	guard	good	deal	concern
asset	volum	suprem	accept	project	win	economist
strong	increas	rate	better	rise	work	invest
chang	know	feder	high	level	banker	confid
demand	reformist	growth	positiv	forward	continu	higher
predict	spread	analyst	export	exposur	brokerag	excess
recov	regul	stronger	bond	treasuri	appreci	strengthen

Appendix B.5 Phase 2 Experiment 1-Examples of unigram features classified to class critical up

Appendix C

revolutionariguard	realexchang	exchangrate	centralbank
supremleader	restructurpropos	inflatexpect	mediumterm
frontiermarket	realeconomi	illegimmigr	estatagent
emergmarket	publicdebt	financiinstitut	debtoffic
supplidemand	loangiven	courtlaw	smallnumber
financiservic	existsharehold	ministrifinanc	jobloss
absolutreturn	debtrisk	economciti	shortterm
productregion	financjob	jobmarket	competittougher

Appendix C.1 Phase 2 Experiment 2-Examples of bigram features classified to class down

convertbond	midmarket	rowprice	absolutreturn
foreigninvestor	assetmanag	investbank	realestat
corporgovern	managteam	legalservic	globalaverag
financicentr	shareholdactiv	productregion	bankmodel
legalinnov	legalsupport	innovlawyer	forumsecur
corporlaw	partnerfreshfield	lawpartnership	governlaw
fixfee	sharetrade	highersalari	sharesale
busidevelop	foundpartner	securlawyer	newmodel

Appendix C.2 Phase 2 Experiment 2-Examples of bigram features classified to class up

southkorea	pensionfund	assetmanag	goldprice
goldmarket	middleast	centralbank	creditmarket
worldgold	bankreserv	goldindustri	housprice
irantrade	britishdiplomat	iranianbank	westernbank
absolutreturn	servicconsult	abudhabi	totalreturn
arabianautomobil	standardbank	ukcommerci	gulfcompani
newlanguag	monthperiod	financjob	hostcountri
goldcouncil	goldportfolio	subprime	financiservic

Appendix C.3 Phase 2 Experiment 2-Examples of bigram features classified to class neutral

householdincom	marketindex	exitstrategi	excesscapac
inflationaririsk	biggestrisk	taxrise	depressgrowth
domesteconomi	consumspend	longterm	recoverislow
fiscaltighten	highertax	remainweak	lossmomentum
mediumterm	oversell	raisedebt	bigdebt
householdconsumpt	bigrisk	surprisevent	householdfirm
employgrowth	capitreduc	sharpfall	weakeconom
boostinventori	publicfinanc	creditcondit	newgovern

Appendix C.4 Phase 2 Experiment 2-Examples of bigram features classified to class critical down

developeconomi	regioninvest	emergeconomi	riserate
globalbond	averagup	helpeconom	strongeconomi
emergmarket	growthstock	debtshare	assetmanag
debtpaid	globalinvest	promotgrowth	stronggovern
overbuy	promotgrowth	investvital	goforward
highinvest	capitaltrust	marketrecovery	highinterest
debtsustain	bankinvest	economiworld	improvmarket
helpeconomi	uptick	risemarket	helppeopl

Appendix C.5 Phase 2 Experiment 2-Examples of bigram features classified to class critical up

Appendix D

INFORMATION and CONSENT FORM

Researcher: Mazen Nabil Elagamy, Staffordshire University, United Kingdom.

Thank you for agreeing to support my research which investigates the use of the Random Forest algorithm to examine critical indicators for stock market movements, illustrated with reference to the Dubai debt standstill. The purpose of the study is to determine whether textual information, such as that contained in articles in the Financial Times, can be extracted and used to identify stock market movements.

This consent form confirms our discussions. I am a PhD student at Staffordshire University and have invited you to take part in this research as a domain advisor, due to your expertise in the financial area. Participation in the study is entirely voluntary and all participants' details will be kept anonymous.

If you would like any details about the outcomes of the research or wish to be involved in any further discussions, I would be happy to provide these. Please contact me through my e-mail: m_elagami@hotmail.com

I confirm that the purpose of the study was fully discussed with me and that I have given my consent to take part in the study:

Name:

Signature:

taking part in this study