



Brunel
University
London

***THE PREDICTIVE POWER OF STOCK MICRO-
BLOGGING SENTIMENT IN FORECASTING
STOCK MARKET BEHAVIOUR***

This thesis is submitted for the degree of Doctor of Philosophy

By

Alya Ali AL-Nasseri

Brunel Business School
Brunel University, London

February 2016

ABSTRACT

Online stock forums have become a vital investing platform on which to publish relevant and valuable user-generated content (UGC) data such as investment recommendations and other stock-related information that allow investors to view the opinions of a large number of users and share-trading ideas. This thesis applies methods from computational linguistics and text-mining techniques to analyse and extract, on a daily basis, sentiments from stock-related micro-blogging messages called “StockTwits”. The primary aim of this research is to provide an understanding of the predictive ability of stock micro-blogging sentiments to forecast future stock price behavioural movements by investigating the various roles played by investor sentiments in determining asset pricing on the stock market.

The empirical analysis in this thesis consists of four main parts based on the predictive power and the role of investor sentiment in the stock market. The first part discusses the findings of the text-mining procedure for extracting and predicting sentiments from stock-related micro-blogging data. The purpose is to provide a comparative textual analysis of different machine learning algorithms for the purpose of selecting the most accurate text-mining techniques for predicting sentiment analysis on StockTwits through the provision of two different applications of feature selection, namely filter and wrapper approaches. The second part of the analysis focuses on investigating the predictive correlations between StockTwits features and the stock market indicators. It aims to examine the explanatory power of StockTwits variables in explaining the dynamic nature of different financial market indicators. The third part of the analysis investigates the role played by noise traders in determining asset prices. The aim is to show that stock returns, volatility and trading volumes are affected by investor sentiment; it also seeks to investigate whether changes in sentiment (bullish or bearish) will have different effects on stock market prices. The fourth part offers an in-depth analysis of some tweet-market relationships which represent an open problem in the empirical literature (e.g. sentiment-return relations and volume-disagreement relations).

The results suggest that StockTwits sentiments exhibit explanatory power in explaining the dynamics of stock prices in the U.S. market. Taking different approaches by combining text-mining techniques with feature selection methods has proved successful in predicting StockTwits sentiments. The applications of the approach presented in this thesis offer real-time investment ideas that may provide investors and their peers with a decision support mechanism. Investor sentiment plays a critical role in determining asset prices in capital markets. Overall, the findings suggest that investor sentiment among noise traders is a priced factor. The findings confirm the existence of asymmetric spillover effects of bullish and bearish sentiments on the stock market. They also suggest that sentiment is a significant factor in explaining stock price behaviour in the capital market and imply the positive role of the stock market in the formation of investor sentiment in stock markets. Furthermore, the research findings demonstrate that disagreement is not only an important factor in determining trading volumes but it is also considered a very significant factor in influencing asset prices and returns in capital markets.

Overall, the findings of the thesis provide empirical evidence that failure to consider the role of investor sentiment in traditional finance theory could lead to an imperfect picture when explaining the behaviour of stock prices in stock markets.

DEDICATION

Dedicated to the loving memory of my late grandmother and grandfather (may Allah grant their eternal peace) who always desired and prayed for my success but did not live to see this great accomplishment. Also, to my parents and family for their unconditional love and support which helped me to achieve my aspiration.

ACKNOWLEDGEMENTS

I am immensely indebted to Allah (The Almighty God) for bestowing on me the knowledge of His creation. I pray to Him for forgiveness, guidance, and assistance and to continually support me towards success in my whole life in this world and hereafter.

I am most grateful for the wise council of my ‘wonderful’ supervisor, Dr. Sergio de Cesare. I have been fortunate and honoured to know him and to work under his supervision. I would like to express my profound gratitude for all his great help support, encouragement, and continuous guidance on this thesis-writing process. My deepest appreciation also goes to Dr. Allan Tucker (Department of Computer Science) for his invaluable guidance, experience, inspiration, support and encouragement throughout my PhD journey. I am particularly grateful to him for working on a couple of joint papers on Text Mining. I owe him my deepest gratitude not only for his very helpful comments and guidance but also for all his kindness to me in boosting my confidence. I feel very lucky to have worked under his guidance and advice. I would also like to express my deepest appreciation to Dr. Faek Menla Ali (Department of Social Science). I am very grateful to him for his continuous support, guidance and invaluable comments and suggestions on the empirical finance part of this thesis. I am very thankful to him for his great advice and valuable comments in writing up and structuring journal papers.

Without the unconditional love and endless support of my parents, I would not have been in a position to complete this PhD research. No words can express my gratitude to them. Thank you, Mum and Dad, for your loving support and endless encouragement. I owe a substantial debt of gratitude and thanks to my lovely sisters, Aisha and Tasnim. Both of you, more than anyone, have always been there for me during the twists and turns and through the ups and downs. I am deeply grateful to you, my gorgeous and amazing sisters, and I can never repay even a little of what you have given me of your exceptional love, constant care and emotional support. My special thanks must also go to my three amazing and wonderful brothers, Hossam, Mohammed and Mansoor, for their invaluable support, care and prayers. All three of you have given me constant support and I truly cannot thank you enough. I love you all.

Special thanks and sincere gratitude are due also to my friends and other family members who have been very supportive during this challenging period of my PhD study.

Finally, I would like to take the opportunity to express my gratitude to members of the StockTwits website for providing me with the relevant data for this research. Without their valuable input, this study would not have been possible.

DECLARATION

I grant powers of discretion to the Librarian of Brunel University to allow this thesis to be copied in whole or in part without the necessity to contact me for permission. This permission covers only single copies made for study purposes subject to the normal conditions of acknowledgment.

PUBLICATIONS AND CONFERENCES

The following journal and conference papers are outputs based on the research conducted during my PhD study:

- Al Nasser, A., Tucker, A. and de Cesare, S., 2015. Quantifying StockTwits semantic terms' trading behavior in financial markets: An effective application of decision tree algorithms. *Expert Systems with Applications*, 42(23), pp.9192-9210.
- Al Nasser, A., Tucker, A. and de Cesare, S., 2014, January. Big Data Analysis of StockTwits to Predict Sentiments in the Stock Market. In *Discovery Science* (pp. 13-24). Springer International Publishing.
- Al Nasser, A., 2014. The Predictive Value of Stock Micro-blogging Sentiments in Predicting Stock Market Behaviour. Proceeding of the British Academy of Management (BAM) Conference. 09-11 September, Belfast Waterfront, Northern Ireland.
- Al Nasser, A., 2014. The Predictive Value of Stock Micro-blogging Sentiments in Predicting Stock Market Behaviour. British Academy of Management (BAM): SIG Workshop, University of East Anglia, Norwich, 12-13 June 2014.
- Al Nasser, A., 2016. Dispersion of Stock Returns and Investor Sentiment: StockTwits Evidence. World Finance Conference. 29-31 July, New York, Manhattan.

Journal Articles Submitted

- "Dispersion of Stock Returns and Investor Sentiment: StockTwits Evidence", submitted to, **Journal of Empirical Finance**.
- "Investors' divergence of opinion and Trading Volume: Evidence from Online Stock Forum", targeted journal, **Journal of Banking and Finance**.

Best Paper Award

- "The Predictive Value of Stock Micro-blogging Sentiments in Predicting Stock Market Behaviour", Paper Presented at Brunel Business School, PhD Doctoral Symposium 2014, Brunel University. (Best Overall Paper Award).

Table of Contents

ABSTRACT	i
DEDICATION	ii
ACKNOWLEDGEMENTS	iii
DECLARATION	iv
PUBLICATIONS AND CONFERENCES	v
CHAPTER ONE: INTRODUCTION	17
1.1 Research Background	17
1.2 Research Problem	20
1.3 Research Aim and Objectives	22
1.4 Research Relevance and Significance	24
1.5 Research Methodology	27
1.6 Thesis Structure	29
CHAPTER TWO: LITERATURE REVIEW	32
2.1 Introduction	32
2.2 Trading Theories	33
2.2.1 Efficient Market Hypothesis	33
2.2.2 Random Walk Theory (RWT)	34
2.2.3 The Conflicting Evidence of the Validity of the EMH.....	35
2.2.4 Theoretical and Empirical Challenges to the EMH	36
2.2.5 From Traditional Finance Approach to Behavioural Finance Theory.....	38
2.3 The Role of Noise Traders in Capital Markets	39
2.3.1 Noise Traders and Stock Price Behaviour	40
2.3.2 Investor Sentiment, Disagreement and Stock Market Relations	42
2.4 Online Stock Forums	44
2.4.1 Internet Message Boards and Financial Markets	44
2.4.2 Financial News Articles and Financial Market.....	49
2.4.3 Micro-blogging Forums and Financial Market.....	52
2.5 Text Mining	54
2.5.1 Text Mining Definition	54

2.5.2 Text Mining Tasks	55
2.5.3 Feature Selection.....	58
2.5.4 Text Mining Techniques	62
2.6 The Role of Classifier in Feature Selection.....	70
2.7 The Literature Gaps	72
2.8 Chapter Summary	75
CHAPTER THREE: CONCEPTUAL FRAMEWORK.....	77
3.1 Introduction.....	77
3.2 Theoretical Foundation	78
3.3 Development of the Conceptual Framework.....	80
3.4 Research Hypothesis	81
3.4.1 Message Volume and Stock Market Features.....	82
3.4.2 Investor Sentiment and Stock Market Features	89
3.4.3 Agreement and Stock Market Features	99
3.5 Chapter Summary	104
CHAPTER 4: RESEARCH METHODOLOGY	105
4.1 Introduction.....	105
4.2 Research Paradigm.....	105
4.2.1 Positivist Paradigm	107
4.2.2 Interpretive Paradigm.....	108
4.3 Research Approach.....	109
4.3.1 Quantitative.....	109
4.3.2 Qualitative.....	110
4.4 Justifications for the Adoption of the Research Approach	111
4.5 Research Design	112
4.6 Data Collection	113
4.6.1 Data Generation Sources.....	113
4.6.2 Instruments.....	116
4.6.3 Statistical Packages	118
4.7 Data Analysis	119
4.7.1 Rationale of Modelling	120
4.7.2 Sampling	121
4.8 Framework Design.....	125

4.8.1 Data Description and Pre-Processing Framework	128
4.8.2 Feature Selection and Construction Framework.....	132
4.8.3 Text-Processing Models.....	138
4.8.4 Performance Evaluation.....	146
4.8.5 Training and Testing	150
4.8.6 Statistical Summary	153
4.9 The Econometric Model	160
4.9.1 The Vector Autoregressive (VAR) Model.....	160
4.9.2 Quantile Regression Approach	162
4.10 Chapter Summary	164
CHAPTER FIVE: TEXT MINING ANALYSIS AND FINDINGS	166
5.1 Introduction.....	166
5.2 StockTwits Sentiment Hand-Labeling.....	167
5.3 Model Building in Weka.....	168
5.4 Feature Selection.....	171
5.4.1 Filter Approach	171
5.4.2 Wrapper Approach.....	175
5.5 Selecting the Best Algorithms	177
5.6 Training and Testing	178
5.6.1 Method (1): Training and Testing Using Automatic Percentage Split (66% Training and 33% Testing) Using One Dataset	178
5.6.2 Method (2): Training and Testing Using Supplied Test Set of Two Separate Datasets (Training set (In-Sample Set) and Testing Set (Hold-Out Set)).....	179
5.7 Overall Classification Distribution.....	180
5.8 Application (1): Application of Wrapper Approach: Bayesian Network Model for Prediction of Investor Sentiment in Capital Market	181
5.8.1 Experiments and Analysis.....	182
5.8.2 Performance Comparison.....	182
5.8.3 Bayesian Network Model for Sentiment Prediction	184
5.8.4 Textual Visualization of features Selection Using Wordle.....	190
5.8.5 Discussion	192

5.9 Application (2): Application of Filter Approach: Quantifying StockTwits Semantic Terms' Trading Behaviour in Financial Markets: An Effective Application of Decision Tree Algorithms	193
5.9.1 System Pipeline.....	194
5.9.2 Trading Strategies Design.....	195
5.9.3 Benchmark Trading Strategies.....	198
5.9.4 Empirical Test and Analysis	200
5.9.5 Performance Evaluation.....	204
5.9.6 Mean-Variance Analysis.....	211
5.9.7 Portfolio Constructions and Investment Hypothesis.....	212
5.9.8 Cumulative Performance of the Sell, Buy and Hold Portfolios.....	213
5.9.9 Investment Hypothesis Evaluation	215
5.9.10 Discussion	220
5.10 Chapter Summary	221
CHAPTER SIX: EMPIRICAL FINANCE ANALYSIS AND DISCUSSION...222	
6.1 Introduction.....	222
6.2 Descriptive Statistics and Preliminary Analysis	223
6.3 Distribution of StockTwits Postings	227
6.3.1 Distribution of StockTwits by DJIA Tickers	227
6.3.2 Distributions of StockTwits by time of day	228
6.3.3 Distribution of StockTwits by Day of the Week	230
6.3.4 Distribution of StockTwits postings over the sample period.....	231
6.4 The Contemporaneous Relationship between Stock Micro-blogging Features and Stock Market Indicators	231
6.4.1 Pairwise Correlation.....	231
6.4.2 Contemporaneous Regression.....	234
6.5 The Lead-Lag Relationship between Stock Micro-blogging Features and Stock Market Indicators.....	238
6.5.1 VAR - Return Model	239
6.5.2 VAR - Trading Volume Model.....	243
6.5.3 VAR - Volatility Model	247
6.6 Impulse Response Functions	251
6.7 Chapter Summary	255

CHAPTER SEVEN: AN EMPIRICAL INVESTIGATION OF THE ROLE OF INVESTOR SENTIMENT IN THE STOCK MARKET	257
7.1 Introduction	257
7.2 The Impact of Investor Sentiment on the Stock Market	258
7.2.1 The Effect of the Change in Investor Sentiment on Stock Return and Volatility (The DSSW (1990) Model)	258
7.2.2 The Effect of Investor Sentiment on Trading Volume	266
7.3 Investor Sentiment Reactions to Different Regimes of the Market (Bull and Bear Markets)	273
7.4 Dispersion of Stock Returns and Investor Sentiment: A Quantile Regression Approach	276
7.4.1 Empirical Methodology and Model Specifications	277
7.4.2 Empirical Results	280
7.5 Investors' divergence of opinion and Trading Volume	292
7.5.1 Empirical Methodology and model specifications	292
7.5.2 Volume portfolio strategies based on disagreement	294
7.5.3 Empirical Findings and Regression Result	297
7.6 Chapter Summary	307
CHAPTER EIGHT: CONCLUSION	308
8.1 Introduction	308
8.2 Research Overview and Key Findings	308
8.3 Research Contributions and Implications	313
8.3.1 Empirical Contributions.....	314
8.3.2 Methodological Contributions	317
8.3.3 Practical Contributions.....	318
8.4 Research Limitations	320
8.4.1 Method limitations	320
8.4.2 Data limitations	321
8.5 Directions for Future Research	322
8.6 Chapter Summary	324
REFERENCES	326
APPENDICES	359
Appendix I	359

Appendix II.....	361
Appendix III	363
Appendix IV	368
Appendix V	371
Appendix VI	380
APPENDICES REFERENCES	383

LIST OF TABLES

Table 1.1: Thesis Structure	29
Table 2.1: Selected studies on Internet Message Boards	48
Table 2.2: A summary of Selected research studies on sentiment polarity.	57
Table 2.3: Biases and assumptions of different classifiers	71
Table 4.1: Ontology, Epistemology and Methodology: differences between Positivist and Interpretive research paradigms	106
Table 4.2: Secondary data collection method: strengths and limitations.....	115
Table 4.3: The Coding scheme for manually-labelled Tweets	125
Table 4.4: The list of the required attributes for Stocktwits collection	129
Table 4.5: Examples of Stocktwits messages	129
Table 4.6: A List of advantages and disadvantages of Naïve Bayes, Decision Tree And Support Vector Machine classifiers	145
Table 4.7: Representation of confusion matrix	147
Table 5.1: The manual classifications of Stocktwits messages	167
Table 5.2: Sample Tweets from training set with manual classification	168
Table 5.3: Summary results of the classification performance evaluation of Nb, Randf and Smo	170
Table 5.4: Features Selected under Filter approach using Information Gain Criteria	172
Table 5.5: The best overall classification accuracy (in %) for The Information Gain subsets for Nb, Randf And Smo.	173
Table 5.6: The average classification accuracy of the “N” best attributes selected under Wrapper method for Nb, Randf And SMO classifiers.....	175
Table 5.7: A comparison of the two Feature Selection (Fs) Methods (Filter And Wrapper) of the same attributes selected for Nb, Randf And Smo	176
Table 5.8: Evaluate Random Forest classifiers on Stocktwits data using % split method.....	179
Table 5.9: Classification accuracy by class using Decision Tree Classifier (Random Forest)	179
Table 5.10: Random Forest classification accuracy of supplied test set and the overall classification distribution	180

Table 5.11: The overall distribution of the total Stocktwits postings of all the 30 companies of the DJIA Index	181
Table 5.12: The experimental results of the feature selection and related average classification accuracy of (Bn, Nb, Randf And Smo) classifiers for all quarters (Qs)	183
Table 5.13: Feature subset selected under Bayes Net Classifier for individual quarters	183
Table 5.14: The decision rules for individual occurrence of the term in the Stocktwits postings	202
Table 5.15: The decision rules for combinations of terms appeared in the Stocktwits postings	204
Table 5.16: The cumulative average returns and the number of sell/buy trades of the tweet term trading TTT strategies	208
Table 5.17 The mean-variance analysis	211
Table 5.18: The term trading strategies in the sell, hold and buy portfolios	212
Table 5.19: Predicting portfolio's trading strategy returns based on the asymmetric effects of the increase and decrease in the mean relative changes of the term related frequencies.	217
Table 6.1: Summary statistics by variable	223
Table 6.2: Summary statistics by company	226
Table 6.3: The list of the DJIA Index stock tickers	227
Table 6.4: Distribution of postings by top ten tickers.....	228
Table 6.5: Pearson Correlation matrix	232
Table 6.6: Contemporaneous Regressions	235
Table 6.7: Result of The Var And Granger Causality tests for stock return.....	241
Table 6.8: Result of the var and Granger Causality tests for trading volume.....	244
Table 6.9: Result of the var and Granger Causality tests for stock return volatility..	250
Table 6.10: Summary of the results of the analysis of Stocktwits features and stock market	256
Table 7.1: The relationship between changes in investor sentiment with stock return and volatility	263
Table 7.2: The relationship between changes in investor sentiment and trading volume	268

Table 7.3: The asymmetric impact of bullish and bearish shift in sentiment on trading volume.....	270
Table 7.4: The asymmetric response of the investor sentiment to the change in stock returns in the bull and bear markets.	276
Table 7.5: Estimated coefficients of the linear OLS and QR models	284
Table 7.6: Testing symmetry of quantile causal effects of the linear model	284
Table 7.7: Estimated coefficients of the non-linear (asymmetric) OLS and QR models	288
Table 7.8: Testing symmetry of quantile causal effects of the non-linear (asymmetric) model.....	291
Table 7.9: Mean portfolio returns by trading volume and disagreement.....	296
Table 7.10: The linear regression model (volume-disagreement)	298
Table 7.11: The non-linear model (asymmetric response of trading volume to the investors' disagreement in the bull and bear market)	301
Table 7.12: Volume Regression.....	306

LIST OF FIGURES

Figure 1.1: Research Design	27
Figure 2.1: Text mining as an interdisciplinary field.....	55
Figure 2.2: a) A simple graphical representation of A Bayesian Network with five nodes and b) A Bayesian classifier where c denotes the class node	64
Figure 2.3: A Decision Tree structure.....	66
Figure 2.4: The separable hyper-plan of Vector Support Machine	68
Figure 3.1: The Conceptual Framework	81
Figure 3.2: The relationship between message volume and stock market variables (trading volume, return and volatility).....	84
Figure 3.3: The relationship between investor sentiment and stock market variables (trading volume, return and volatility).....	90
Figure 3.4: The relationship between investor disagreement and stock market variables (trading volume, return and volatility).	100
Figure 4.1: Framework Design	127
Figure 4.2: The feature selection process	134
Figure 4.3: The process of filter feature selection	135
Figure 4.4: The process of wrapper feature selection	137
Figure 4.5: The maximum hyper-plan of support vector machine	142
Figure 4.6: The soft margin loss and ϵ -insensitive loss function for a linear SVM...	144
Figure 4.7: Training and testing procedure to assess the model accuracy.....	151
Figure 4.8: The relationship between Stocktwits features and stock market indicators	153
Figure 5.1: Comparative performance of nb, randf and smo classifiers	171
Figure 5.2: The overall classification accuracy for the “best ranked” attributes by IG criteria.	174
Figure 5.3: Results of an extracted Bayesian Networks model of buy sentiment	185
Figure 5.4: Results of an extracted Bayesian Networks model of sell sentiment.....	187
Figure 5.5: Results of an extracted Bayesian Networks model of hold sentiment	189
Figure 5.6: The conditional probability distribution of the most common words related with the (a) buy, (b) sell and (c) hold sentiment	190
Figure 5.7: The textual visualization of feature selections of Bayes Net for buy, sell and hold sentiment over four quarters.	191

Figure 5.8: System Pipeline	195
Figure 5.9: The standard deviation of 1,000 simulations of average returns using purely Random Investment Strategy.....	199
Figure 5.10: An extracted version of the visualised Decision Tree Model	201
Figure 5.11: A comparison of the monthly average cumulative performances of the trading strategy of the tweet terms “report”, “support”, “report+intc” and “break+support” with the Random Investment Strategy.	206
Figure 5.12: Performance of TTT investment strategies based on term related frequency.....	211
Figure 5.13: Performance of sell, buy and hold portfolios strategies.	214
Figure 6.1: The monthly average movement of stocktwits variables (bullishness, message volume and agreement) and financial market variables (trading volume, returns and volatility).	225
Figure 6.2: The hourly distribution of Stocktwits.....	229
Figure 6.3: The distribution of Stocktwits posts throughout the week	230
Figure 6.4: The daily Stocktwits messages.....	231
Figure 6.5a: Generalised impulse response functions of short-run Granger Causality between stock return and Stocktwits variables.	253
Figure 6.5b: Generalised impulse response functions of short-run Granger Causality between trading volume and Stocktwits variables.....	254
Figure 6.5c: Generalised impulse response functions of short-run Granger Causality between trading volume and Stocktwits variables.....	255
Figure 7.1: The impact of noise trader sentiment on stock returns and volatility	259
Figure 7.2: Estimates of the linear OLS and QR models.....	280
Figure 7.3: Estimates of the nonlinear (asymmetric)OLS and QR models.	287

CHAPTER ONE: INTRODUCTION

1.1 Research Background

The advance of the World Wide Web (WWW) has generated a series of changes in the business environment. The Web is changing the way businesses are running and performing. Before the emergence of the Internet, companies disclosed information on their performance or other aspects through various media such as earnings reports, corporate communications and management interviews (Weston, 2001), which took a long time to disseminate among interested parties. The investor base is constantly on the lookout for any new information from such events that may help them increase their returns or reduce their risk exposure (Schillhofer, 2008). On the other hand, the companies and markets also took a long time to dispel market rumours and false information.

The innovations of Web 2.0 technology and social media have resulted in even more progressive changes and are characterised by rapid information dissemination as well as retrieval (Ellison and Nicole, 2007). Investors have been dramatically affected by these changes. The Internet has made vast amounts of information available to investors and stakeholders, altering the way information is gathered and exchanged as well as the way investors deal with and act upon that information (Barber and Odean, 2001). Any information (good or bad) about a particular company (e.g. product, service, person etc.) can be disclosed at the click of a mouse (Acemoglu et al., 2010; Brown and Duguid, 2002) or through micro-blogging services such as Twitter. A considerable amount of literature has been published on the use of Twitter feeds. These studies have made use of Twitter posts to predict various phenomena such as box office revenues (Asur and Huberman, 2010) and the spread of swine flu (Ritterman et al, 2009) and disaster news (Doan et al., 2012). Recently, scholars have also addressed the extraction of sentiments from Twitter feeds by investigating the relationship between sentiments extracted and financial market variables. Their findings reveal that sentiments play a critical role in predicting the short-term financial performance of financial securities and assets pricing (Qiu et al., 2011).

Chapter One: Introduction

Over the past few decades, a large and growing body of literature has provided empirical evidence that investor sentiment is closely associated with stock price (i.e. Baker and Wurgler, 2006, 2007; Brown and Cliff 2004, 2005; Verma Verma, 2007; Lee et al., 2002). These studies suggest that the issue now facing financial economists is not only the predictive ability of investor sentiments to influence security prices but also the extent to which investor sentiment can impact the stock market. While news undoubtedly influences security prices in the stock market, public mood and emotions (sentiments) may play an equally important role (Bollen et al., 2011). Studies that investigate the impact of investor sentiment on the stock market rest on critical examinations of the assumption underlying behavioural finance theory. Psychological research has conclusively demonstrated that emotions as well as information play a significant role in altering human decisions (Kahneman et al., 1979). Through the role played by noise traders in determining security prices (De Long et al., 1990), behavioural finance has provided further evidence that investment decisions are significantly driven by emotions and sentiment (Nofsinger, 2005). It has been validated that the market is completely driven by sentiments and the bullishness of investors' decisions (Qian and Rasheed, 2007). This view is supported by most recent studies, which have found a fruitful area of research to investigate the effect of public/investor sentiments in predicting stock price movements (Oh and Sheng, 2011; Sprenger et al., 2014; Bollen et al., 2011).

One popular area in financial analysis and computational finance for pattern recognition and machine learning applications is the instant access to news on companies and the highly dynamic and data-intensive capital markets. In recent years, there has been an increasing interest in stock market predictions using various statistical tools and machine learning techniques. Different methodologies have been developed with the aim of predicting the direction of security prices as accurately as possible (Guresen et al., 2011). This is still an on-going field of extensive research; however, no methods have yet been discovered and proved capable of undertaking such a task. In spite of the continuing efforts by researchers to solve this issue, results have been inconclusive and few successes have been achieved. Moreover, several studies investigating stock market prediction have obtained results that are in line with the widely accepted theories, implying the difficulty of predicting the price of a security and suggesting that it is an impossible task (Butler and Malaikah, 1992;

Chapter One: Introduction

Kavussanos and Dockery, 2001; Gallagher and Taylor, 2002; Qian and Rasheed, 2007).

Two widely accepted theories are invoked when the following question is raised: Can stock prices truly be predicted? One such theory is the Random Walk hypothesis (Malkiel, 1996), which states that the price will follow a random pattern and suggests that attempts to predict the stock market will never be accurate since prices are randomly determined. Furthermore, the efficient market hypothesis (EMH) (Fama, 1965b) states that market prices reflect all publicly available information and that everyone has same degree of access to that information; hence, the financial market is said to be “informationally efficient”. Therefore, these theories suggest that attempts to predict market values are based solely on chance and that stocks are traded at their fair values, making it impossible for investors either to purchase undervalued securities or sell stocks at higher market values; it is therefore impossible to beat the market (Xu, 2012).

Recent developments in Information Technology (IT) have heightened the need for the adoption of various social media platforms as vital communication tools in the business world. Most companies around the world acknowledge this importance and have started to utilise social media platforms as further communication tools to keep their stockholders empowered and informed. New media channels such as Virtual Investing Communities (VIC) and financial blogs (such as Yahoo Finance, Seeking Alpha and StockTwits) publish relevant and valuable user-generated content (UGC) and data (e.g., investment recommendations). These media allow investors to view the opinions of a large number of users as well as share and exchange investment ideas. UGC enables investors to take a more active role as capital market players and reach (and be reached by) almost everyone, anywhere, anytime. By allowing investors to monitor the thoughts and opinions of other investors on specific securities of interest to them, it may be possible to improve and enhance their ability to make better informed investment decisions with the potential for greater returns on their investments. There is, therefore, a very appealing challenge to researchers to explore how investors react and interact in such VICs and to investigate whether they help in predicting future stock price movements in capital markets.

Chapter One: Introduction

Data mining and sentiment analysis are the techniques that have been adopted most recently by researchers. The generalisability of much published research on this issue has produced interesting results (Bollen et al., 2010; Zhang, 2009; Antweiler and Frank, 2004a and b; Das and Chen, 2007; Sprenger et al., 2014). Sentiment analysis and text mining methods function to extract meaning and knowledge information from various sources on the Web including company websites, social networking sites and micro-blogs. Researchers, in their analysis of textual data, have achieved great success in predicting stock market prices using text mining and machine learning applications. However, despite a widespread belief that sentiments from investment community forums have the power to predict financial market trends, little is yet known about whether these sentiments play any significant role in predicting stock price movements (Tumarkin and Whitelaw, 2001; Das and Chen, 2007; Antweiler and Frank, 2004a and b; Sprenger et al., 2014; Bollen et al., 2011; Zhang, 2009). This thesis attempts to study the predictive power of collective sentiments of stock micro-blogging websites on the stock market. Stocktwits.com (<http://www.stocktwits.com>), with its leading social network and financial community forum, provides real-time investment ideas with a high volume of stock message postings. This thesis bridges this research gap by hypothesising that stock micro-blogging forums may enable us to observe previously unavailable aspects of the dissemination of financial information in online investment community channels and their predictive value and to determine whether or not they have the power to affect or alert investors' decision making toward specific types of traded stocks in the financial market.

1.2 Research Problem

Stock micro-blogging is considered a new topic that has little been addressed by scholarly research. Many research studies have investigated the relationship between stock trading message boards and the financial market. These studies include Wysocki (1998), Koski et al. (2004), Antweiler and Frank (2002, 2004a and b), Tumarkin and Whitelaw (2001) and Jones (2006). While this stream of research has focused on and is limited to quantitative data (for example, message volume (Wysocki, 1998; Jones, 2006) and users' ratings (Tumarkin and Whitelaw, 2001)), a study conducted by Antweiler and Frank (2004a) has focused on qualitative as well as

Chapter One: Introduction

quantitative data analysis of the Internet message boards posted on Yahoo Finance and Raging Bull to determine the correlation between the activity on the Internet message board and stock volatility and trading volume. Gu et al., (2006) adopted a different approach to address the relationship between stock message boards and capital markets and focused on the most important element in the capital market (investors). They used sentiment analysis on the stock message board in Yahoo! Finance to examine its effect on investors' decisions. Their study examined the predictive power of the stock message board on the future abnormal return. Each message was classified as either positive or negative and was then analysed as a 'buy' or 'sell' decision. Their analysis is based on a trading strategy that follows this pattern: selling stock with high sentiments while buying stock with low sentiments.

Another stream of studies focused on using financial news articles to predict stock market movements (Schumaker and Chen, 2009; Gidofalvi, 2001; Chen et al., 2012). Chen et al. (2012) examined peer-based advice, which is transmitted through social media, to determine whether it plays a role in the financial market. This study employed a textual analysis of journal articles published on Seeking Alpha, which is among the most popular social media platforms among investors. They argued that most of the views that are expressed in these articles are strongly associated with subsequent stock returns and also help in predicting earnings surprises.

As micro-blogging has appeared relatively recently, only a few research works have been devoted to this topic. Previous research studies have focused on both financial news articles and Internet stock message boards in predicting the financial market. However, little is known about the effect of stock micro-blogging contents on the stock market. Despite the similarity of these established financial blogging forums, the unique characteristics of stock micro-blogging websites make it difficult to generalise the results of previous studies on stock message boards and financial news articles for the following reasons. Although micro-blogging services provide an easy way of sharing status messages either publicly or on a social network, the unique features of simple messages of 140 characters in Twitter posts is considered a brief form of information that investors and other users can easily read and follow, unlike lengthy financial news articles containing unlimited word counts that may cause investors to ignore them. In addition, shorter posts take up less of the user's time and increase the volume of postings (Java et al., 2007). Another reason is that real-time

Chapter One: Introduction

conversations on stock micro-blogging forums produce up-to-date information on all stocks compared to separate bulletin boards for each individual company, which result in out-dated information if no new posts have recently been entered on those boards. Furthermore, real-time messages posted just as an event occurs are considered a major contributory factor in the popularity of micro-blogging (Claburn, 2009). Moreover, since financial information such as annual reports, earnings announcements and company press releases are infrequently produced for the public, the real-time streaming of micro-blogs provides new information that is frequently available to investors. Due to the value-weighted information produced in stock micro-blogging forums as well as the limited nature (140-character message posts) compared to stock message boards and news article forums, researchers still need to address these issues to investigate whether these distinct and unique characteristics are effective in different contexts such as predicting security price movements in the capital market.

In this thesis, the issue of stock market prediction is combined with a complex text-mining task. Therefore, the author will attempt to solve this problem by using different classifier techniques of machine learning algorithms for the purpose of mining the raw StockTwits posts and transforming them into linguistic textual representations such as the 'bag of words'. The nature of the contents of such posts (such as the use of abbreviations, emoticons and poor grammar) presents a difficult task for natural language processing. Thus, to overcome these problems and to reduce the dimensions of the raw data, different filtering methods will be applied to finally approach the overall aim of predicting stock price movements by building a regression model using different statistical and analytical approaches.

1.3 Research Aim and Objectives

"Communities of active investors and day traders who are sharing opinions and in some case sophisticated research about stocks, bonds and other financial instruments will actually have the power to move share prices ...making Twitter-based input as important as any other data to the stock"

TIME (2009)

Chapter One: Introduction

The purpose of this thesis is to propose new directions in the roles of investor sentiment that researchers might implement in the analysis of the explanatory power of stock micro-blogging sentiments for future stock price behaviour. The ultimate goal is not to build an ideal model for stock market prediction but to explore whether and to what extent stock micro-blogging affects stock market directions and how it helps to predict the financial market. This study is in line with previous research by Antweiler and Frank (2004 a and b), Oh and Sheng (2011) and Sprenger et al. (2014). The purpose of this study to use sentiment analysis (Das and Chen 2007; Tetlock, 2007; Bollen et al., 2011) and a predictive analytics (Shmueli and Koppius 2011) approach to understand the predictive relationship between the most important market features, such as market return, volatility and trading volume with corresponding stock micro-blogging sentiments (e.g. bullishness, message volume and level of agreement among messages). In this thesis the term “predicting” provides means of anticipating and/or forecasting future actions and behavior of price movements or investors’ behavior in stock market.

The primary aim of this research is, therefore, *to provide an understanding of the predictive value of Stock micro-blogging sentiments in forecasting future stock price directional movement and future stock market performance while determining the most suitable and accurate text mining techniques for sentiment analysis.*

Accordingly, four research questions will be investigated in order to achieve the aim of this thesis:

- 1- To what extent can stock micro-blogging features (e.g. bullishness, message volume and level of agreement) predict stock market behaviour (return, volatility and trading volume)?
- 2- What role does investor sentiment play in determining assets return, volatility and trading volumes in the capital market?
- 3- Is disagreement among messages associated with more trades?
- 4- Can text-mining techniques accurately predict sentiment analysis on StockTwits? And how effective are feature selection methods in improving the sentiment classification accuracy of these techniques?

In order to address the research aim and attempt to answer the research questions, the following research objectives will be met:

Chapter One: Introduction

- To conduct a critical literature review in the area of online investment forums and their effect in predicting financial market movements in general, with particular emphasis on a stock micro-blogging website (StockTwits) in order to examine the predictive power of stock micro-blogging sentiments in forecasting stock market prices.
- To propose a sentiment analysis and predictive analysis approach to understand the predictive correlation between stock micro-blogging sentiments and stock market returns, volatility and trading volume.
- To identify which textual analysis and data mining techniques are more appropriate for sentiment analysis in stock micro-blogs.
- To evaluate and investigate three selected classification algorithms (Naïve Bayes, Decision Tree and Support Vector Machines) based on Weka.
- To investigate the relationship between StockTwits features and the most important stock market features (stock return, volatility and trading volume) using the more suitable statistical and analytical approaches.
- To explore and analyse whether collective users' sentiments on StockTwits have predictive power in forecasting investors' trading decisions through different applications of machine learning techniques.
- To provide an in-depth analysis exploring some of the relationships that are most intriguing and supported by the resulting empirical evidence set out in the forthcoming analysis and findings chapters of the thesis.

Having discussed the aim and the objectives of this research, in the next section the author discusses the importance and significance of the research study.

1.4 Research Relevance and Significance

The continuous growth of digital-based information in financial markets had led to the identification of various issues associated with handling massive amounts of information. In all aspects of life, continuing growth in the quantity of information becomes a very difficult task to handle manually (Indurkha et al., 2005). In any corporate business environment, the financial information is reported in the form of unstructured data, i.e. documents (such as annual reports and earnings announcement reports), web pages (Intranet and the Internet) and social networking sites. The

Chapter One: Introduction

majority of companies' estimated information is in the form of textual information such as emails and memos (Tan, 1999; Karanikas and Theodoulidis, 2002). Finding an effective way of handling and carefully analysing these sources of information could provide a competitive advantage to a company, leading to successful contributions in the area of the knowledge-based economy. Therefore, the need for automated methods to handle such a large quantity of textual information has become significant and necessary (Lagus, 2000).

The field of Text Mining (TM) has become extremely important for solving these issues of managing and mining large textual databases using automated methods and algorithms (Spinakis and Peristera, 2004; Fan et al., 2006). One of the most widely applied functions of text mining techniques is the application of future predictions to anticipate trends based on time-dependent data while associating these patterns with other patterns extracted from the data under analysis. The relevance and importance of this research study has stemmed from the need to use automated textual analysis techniques to extract sentiments to predict future stock price behaviour in capital markets. In this thesis, attempts have been made to predict stock market price patterns (such as returns, volatility and trading volume) to test their associations with other patterns extracted from a stock micro-blogging forum (StockTwits), such as sentiments, message volume and level of agreements.

Stock micro-blogging services such as StockTwits have become very popular communication tools among various investment community platforms. Millions of people share opinions and thoughts on different traded securities in the stock market. Therefore, StockTwits is a rich source of financial data for opinion mining and sentiment analysis. As it is a relatively new platform in the virtual investing community, few research studies have addressed this topic. While a few empirical studies (Sprenger et al., 2014; Rao and Srivastava, 2014; Oh and Sheng, 2011) have validated the predictive power of stock micro-blogging sentiments, this issue still requires in-depth investigation to provide a greater understanding of whether discussions in stock micro-blogging forums are leading the movement in the stock market and whether the investment ideas shared in such forums can or should be used by investors to make their investment choices or decisions. Therefore, this study is considered to be relevant and timely considering the boom in financial blogs and online investment community forums.

Chapter One: Introduction

Stock market predictions coupled with complex data mining techniques represent an on-going field of research that has captured the attention and interest of both researchers and practitioners. It is among the top critical issues on the agendas of today's financial analysts and investment advisors who are seeking to provide better investment ideas to their investors and the best stock price predictions in financial markets. This study contributes to a deeper understanding of the predictive ability of micro-blogging services such as StockTwits sentiments in forecasting stock market movements. The value of this research is to be found in the practical contribution to both companies and their related investors by providing accurate stock price predictions of the securities traded and discussed in stock forums; as a result, investors might make better investment decisions and receive higher returns on their investments.

Researchers and practitioners alike find it a fruitful area of research to pay attention to the boom in financial blogs and online investment communities among financial professionals, investment analysts, investors and their peers (Antweiler and Frank, 2004b; Business Week, 2009). More than ever before, and with the massive amount of information available to investors through various virtual investment community platforms, greater incentives have been given to financial researchers to address and understand the way in which information is produced and spread in the new media channels and how this might affect stock market predictions and the decision-making process. It has become relatively easy and less costly for any interested investors to receive a free form of information (such as investment advice and prediction opinions on particular security prices) that has been analysed by professional and non-professional analysts due to their active and interactive participation in social media channels (Saxton, 2012). In fact, the new forms of media rely heavily on participant-generated content and bottom-up knowledge creation (e.g. O'Reilly, 2007). Participants who actively engage in stock market predictions may make better predictions than spammers (noise traders) who simply make noise through random guesses. Investors pay much attention to these analysts' predictions and base their investment decisions on them. Therefore, it is very challenging for companies to monitor the types of information that analysts provide to investors as this may greatly affect their corporate profitability and image.

1.5 Research Methodology

This research employed a positivist approach to design the research problem and applies quantitative means to collect and analyse the data. To assess in conducting the study appropriately and to help answering the research questions to get the most valid findings, a proper research design therefore must be needed. The research design is based on a research model that is perceived as a sequence of interrelated stages, where the next stage cannot be achieved successfully without the accomplishment of the preceding stage (Sarantakos, 2005). The successful achievement of these sequence stages will result in the research questions being answered. Figure 1.1 lists the stage-by-stage process that is used to conduct the study.

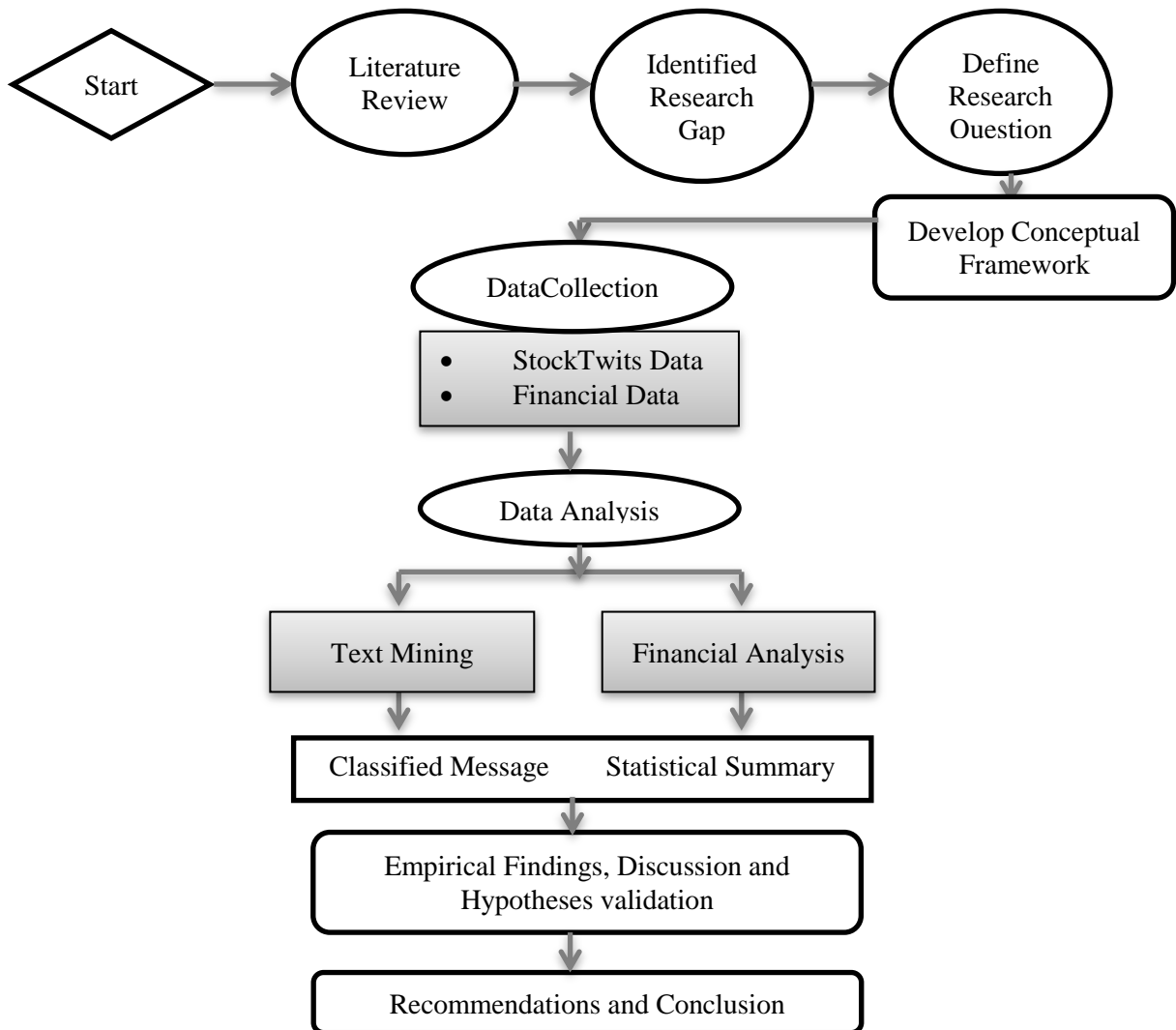


Figure 1.1: Research Design

Chapter One: Introduction

The research design depicted in Figure 1.1 shows that the hypothetico-deductive approach is adopted in this research study, which will enable the researcher to answer the research questions while providing a justification of the hypotheses. In this research process, the initial step began with a critical review of the literature and previous empirical studies in order to provide insights for an understanding of the research field while helping the researcher to identify the research gap in the literature. Formulating and defining the research questions are the next stage that researcher might be doing after identifying the gap in literature. Once the research gap was identified and the research questions are defined, a conceptual framework was developed to provide indications of how this research might be empirically conducted. In the conceptual framework, several features from both StockTwits (e.g. sentiment and message volume and level of agreement) and the stock market (e.g. return, volatility and trading volumes) have been connected regarding an understanding of the predictive power of stock micro-blogging sentiments in predicting stock price behaviour. It also explains the associations and relationships between those feature variables. Then, to test the model developed in the conceptual framework, data must be collected to validate the research hypotheses. Two sources for data collection are used for this thesis: StockTwits data and financial data.

It can be seen, from the above discussion of the sources of data collection, that this research is based on the positivist philosophical approach. The initial stage in any positivist approach is to conduct a literature review and develop a conceptual framework to facilitate hypotheses testing (Cohen et al., 2000). Therefore, in this research two main sources of secondary data can be collected from two fields: an online stock forum (Stocktwits) and stock market data (daily stock prices). Different text-mining techniques and computational finance approaches are used for further analysis of the collected data in order to validate the research hypotheses. This research study encompasses two popular areas of research for pattern recognition: text mining and financial analysis. This research study focuses on predicting stock market dynamics using various statistical tools and machine learning techniques. Different text-mining techniques will be used to perform the textual analysis of StockTwits data to classify messages and construct the StockTwits variables. Then, various statistical techniques and econometrics modelling will be employed to test the correlations with stock market variables; the empirical findings will then be discussed to validate the

Chapter One: Introduction

existing hypotheses. The conclusion of this study provides a broad discussion of the findings, recommendations for future research and the study's limitations.

1.6 Thesis Structure

The thesis is made up of seven chapters, as shown in Table 1.1, and is organised according to the recommendation of Phillips and Pugh's (2005) seminal book *How to get a PhD*.

Table 1.1: Thesis Structure

Background Theory	Chapter 1	Introduction
Focal Theory	Chapter 2	Literature Review
Data Theory	Chapter 3	Conceptual Framework of the predictive power of Stock Micro-blogging in predicting stock market movements
Novel Contribution	Chapter 4	Research Methodology
	Chapter 5	Text-Mining Analysis and Results
	Chapter 6	Empirical Finance Analysis and Discussion
	Chapter 7	Empirical Investigation of the Role of Investor Sentiments in Stock Market
	Chapter 8	Conclusion

Source: Adopted with modification from Phillips and Pugh (2005)

Chapter one (this chapter) is an introduction to the research, where the thesis argumentation is presented. This chapter provides the research background and outlines the broad field of the study. The aim of this chapter is to orientate the reader by providing an overall picture of the rest of the thesis. The chapter includes a brief description of the research background and a formulation of the research problem. The research aim and objectives and a brief explanation of the research design are all provided in this chapter, with a justification of the relevance and significance of the research undertaken. A structural outline of the thesis is given at the end of this chapter.

Chapter One: Introduction

Chapter Two aims to build a theoretical foundation for the research by critically reviewing existing state-of-the-art literature. This chapter reviews the existing literature on two fields of study: text mining and empirical finance. The examination of these fields establishes the boundaries and identifies gaps in existing research. It collects and consolidates relevant literature on related trading theories that serve as a theoretical base of this research study. It also points out the significant role played by noise traders in the capital market while examining the behaviour of the stock market in relation to noise traders' sentiments. This chapter extensively reviews the various effects of different online investment forums (stock message boards, financial news articles and stock micro-blog forums) in predicting the financial market. It then addresses the text mining techniques and sentiment analysis and identifies the usage of those techniques in performing a textual analysis of online discussion forums.

Chapter Three provides the proposed research's conceptual framework for the predictive value of stock micro-blogs in predicting stock market movements. The construct of this chapter is to develop research hypotheses to examine the predictive power of stock micro-blogging features in forecasting future stock price behaviour in the capital market. To address the hypotheses effectively, the researcher extensively reviews relevant subject areas such as market behaviour; this leads to the amplification and clarification of the research area and the development of the theoretical/conceptual perspective of the research.

Chapter Four presents and identifies the chosen research paradigm and outlines the methodology used in the research. It discusses the research approach and methods applied to conduct the empirical investigation of the research while rationally justifying the selection of particular research methods. This chapter also explains in detail the data collection method and highlights the data analysis, statistical techniques and the framework design adopted to carry out the empirical investigation of the research. Additionally, this chapter addresses the kind of data and the appropriate methodologies required to extract and examine each of the variables.

Chapter Five reports the key findings and analysis of the performed text-mining technique to predict sentiments from online financial text using StockTwits postings. This chapter first offers a comparative investigation of classification performances of different machine learning algorithms while exploring the

Chapter One: Introduction

effectiveness of feature selection methods in improving the sentiment classification accuracy of each classifier algorithm. It then presents two different applications utilising the filter and wrapper approaches to feature selection in predicting investor sentiment decisions. This has practical implications, providing investors and their relevant peers with a decision support mechanism in the financial market.

Chapter Six discusses the key findings and analysis of the predictive relationship between StockTwits features and financial market indicators. This chapter first presents the preliminary findings of the data analysis as well as the descriptive statistics of the variables examined in this thesis. It then investigates the contemporaneous and lead-lag relationships between the StockTwits variables and other market indicators by employing a Vector Auto Regressive (VAR) model to provide an answer to the research question that investigates the extent to which StockTwits features might potentially explain the financial market variables.

Chapter Seven reports the main findings of the empirical investigation of the role played by investor sentiment in determining asset-pricing behaviour in the capital market. In this chapter, this thesis focuses on the impact of investor sentiments on stock returns, volatility and trading volume. It also examines the non-linear relationship between StockTwits variables and financial market dynamics to empirically investigate the effect of asymmetrical behaviour of investor sentiment by distinguishing between bullish and bearish sentiments on stock returns. This chapter also captures the asymmetry in the predictive power of investors' disagreement in trading volume in two different regimes/patterns of return: bull and bear markets.

Chapter Eight is the final chapter of the thesis, providing an overall summary of the work. It summarises the main results and findings of the research and highlights its theoretical, methodological and practical contributions. Finally, it provides suggestions for future research directions.

CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction

The previous chapter introduced the research background and overall structure of this study. The primary purpose of this chapter is to explore in depth the related literature on online investment forums and the effect of such forums on financial and stock markets. The literature review provided in this chapter is divided into three parts. The first part discusses the theories utilised as a foundation for this research. The second part provides a general background to the role played by noise traders in determining assets pricing in capital markets. The chapter then offers a critical review of research relating to different types of online stock forums while highlighting their effect on financial markets. The third part of this chapter discusses different text mining tools and techniques used to classify various types of text (e.g., tweet messages in this research study). This part also highlights the underlying role played by text mining in online stock forums and financial markets. In doing so, the chapter presents the theoretical background to provide a reflective and comprehensive understanding, which is used in crafting the conceptual research framework presented in the next chapter.

The literature referenced in this chapter offers solid evidence of the predictive ability of stock micro-blogging forums in forecasting stock market behaviour. The scope of this thesis is to investigate the effective power of stock micro-blogging features in predicting different financial market indicators.

The chapter is divided into eight different sections, including this introduction, as follows. Section 2.2 discusses the trading theories that provide theoretical justification for the claims of this thesis regarding the predictive power of stock micro-blogging sentiments in forecasting stock price movement in stock markets; these theories are utilised as a foundation for this research. Section 2.3 reviews the various roles played by noise traders in determining asset prices in capital markets. Section 2.4 discusses the three main online stock forums and highlights the underlying effect of each of these forums in predicting various financial market indicators. Section 2.5 defines text mining and identifies its various tools and

Chapter Two: Literature Review

techniques while further emphasising the feature selection process for classification problems. Section 2.6 highlights the roles of different classifiers in feature selection. The research gap is identified and discussed in section 2.7. Finally, section 2.8 provides a brief summary of this chapter.

2.2 Trading Theories

Stock market efficiency constitutes one of the most exciting fields in finance to have emerged since the 1960s. A considerable amount of academic research has been published and reported in finance journals and conferences in the past decades on whether stock markets are efficient and to what extent stock prices can be predicted. To address the topic of future stock price prediction, several theories can be considered relevant. Several works have attempted to investigate the predictability of stock prices while providing an answer to the common question: Can stock prices really be predicted? There are two main theories: 1) Efficient Market Hypothesis (EMH) and 2) Random Walk Theory (RWT). The next subsection briefly discusses the relevance of EMH and RWT while highlighting the major theoretical and empirical challenges that question the validity of EMH and RWT.

2.2.1 Efficient Market Hypothesis

EMH was and still is one of the leading theories of traditional finance. It is the most widely accepted theory among financial economists (Malkiel, 2003). The idea of market efficiency was originated by Eugene Fama in the 1960s (Fama, 1965a). The EMH states that the market prices reflect all publicly available information and that everyone has same degree of accessibility to that information; hence, the financial market is said to be “informationally efficient”. This implies that average investors, both individual and institutional, cannot consistently beat the market (Xu, 2012). Therefore, it is impossible for investors to achieve a return above the average risk-adjusted return using investment strategies based on publicly available information (Fama, 1970, 1991). This suggests that the stocks are traded at their fair value and investors are advised to passively buy and hold the investment portfolio rather than engage in active investment strategies because security prices are extremely efficient and are always priced correctly.

Chapter Two: Literature Review

There are three major versions of EMH: weak, semi-strong and strong forms. The weak form of EMH assumes that the price of the traded security is reflected only by current publicly available information in the market. It argues that past price and volume are independent and cannot predict the future price direction of the traded security. The semi-strong version is the most widely accepted belief that goes a step further by integrating all publicly available information (current and past information) and instantly embedded new public information in the current price of the traded assets (Malkiel, 2003; Sprenger et al., 2014). The semi-strong form of EMH also incorporates the weak form hypothesis. The strong version of the EMH reflects all information both public and private information while incorporating both the weak form and semi-strong form of EMH. It also reflects the hidden information such as insider information in the share price in addition to past and current publicly available information. It contends that no investor would be able to gain excess returns above the average investor even if he/she was given new information. It assumes a perfect market and concludes that it is impossible to consistently outperform the market (Schumaker et al., 2011).

2.2.2 Random Walk Theory (RWT)

The efficient market hypothesis is closely associated with a terminology heavily discussed in the finance literature: “random walk” (Malkiel, 2003). The random walk idea states that the price will follow a random pattern due to the random arrival of new information. It assumes that all subsequent price changes represent random departures from previous prices. It contends that when the information is directly reflected in the stock prices, the past changes in prices and historical news have no relationship with the future price change; rather, it will only reflect future news. Malkiel (1973) suggests that the random walk theory provides a different perspective in stock price prediction. According to the random walk theory, it is believed that predicting the stock market is impossible since prices are randomly determined. While this theory assumes equal accessibility of publicly available information to everyone on the market, it is therefore similar in its theoretical perspective to the semi-strong version of the EMH; however, this theory confirms that the prediction is still impossible even when such information has become available (Schumaker et al., 2011).

Chapter Two: Literature Review

2.2.3 The Conflicting Evidence of the Validity of the EMH

The EMH was theorised in the 1960s and was generally accepted until the 1990s. After the “crash of 1987”, the “dot bubble” and the “subprime mortgage crisis”, the validity of EMH has become a matter of great concern for financial economists, academics, investors and analysts alike. Numerous research studies have questioned the validity of EMH and RWT in predicting stock markets. Both EMH and RWT have been tested extensively and there is a huge divergence of results relating to their validity. These results are mixed and sometimes contradictory (Qian and Rasheed, 2007). An early body of research advocated and supported both theories (Cootner, 1964; Jensen, 1978; Lo and MacKinlay, 1988) and was unable to reject the market efficiency and the random walk behaviour of stock prices. Many financial economists agreed with Jensen’s (1978) view that "there is no other proposition in economics which has more solid empirical evidence supporting it than the Efficient Markets Hypothesis."

Nevertheless, some financial academics claim that the EMH and RWT are flawed concepts and are no longer valid in the post-financial crises era. For example, a large and growing body of literature has contradicted the EMH and RWT and questioned the EMH’s basic assumption that stock market prices follow a random behaviour (Butler and Malaikah, 1992; Kavussanos and Dockery, 2001; Gallagher and Taylor, 2002; Qian and Rasheed, 2007). Fama (1991) has even stated that the extreme version of efficient market hypothesis is surely false. A supportive argument in rejecting the EMH is provided by Lo and MacKinlay (2011), who demonstrated that the past prices could be used to provide some level of prediction of future return.

In the context of online financial forums, recent studies have proved that stock market prices do not follow a random walk pattern and that some level of prediction may be possible (Bollen et al., 2011; Gilbert and Karahalios, 2010; Sprenger et al., 2014; Mao et al., 2011). For example, Bollen et al. (2011) argue that since the time occurrence of the news is unpredictable, very early indicators can be extracted from online forums, such as blogs and twitter feeds, to predict changes in various economic and commercial indicators. Their study proved that such a case of news unpredictability could conceivably be the same for the stock market. They showed how the collective mood on Twitter (positive and negative tweets) helps to predict movement on the Dow Jones Industrial Average (DJIA). There are many examples

Chapter Two: Literature Review

showing how online chat forums can be used to predict economic and commercial indicators. For example Gruhl et al. (2005) studied how online discussions can predict book sales and influence customers' purchase decisions, while Mishne and Glance (2006) and Asur and Huberman (2010) investigated whether movie sales figures and box office movie revenues can be accurately predicted.

This discussion of the broad literature on testing the validity of the Efficient Market Hypothesis and the Random Walk Theory appears to have revealed mixed results. In general, much of the early work tends to support the random walk model, while the majority of the studies conducted recently shows contradictory results and have proved that stock prices do not completely follow random walk behaviour and that some degree of prediction is possible. It is, therefore, important to shed light on the theoretical and empirical challenges to the EMH, which will be discussed in the following section.

2.2.4 Theoretical and Empirical Challenges to the EMH

Despite being empirically and theoretically accepted beliefs among financial economists, the EMH assumptions have been called into question; they face both theoretical and empirical challenges and have gradually lost ground. The first challenge to the EMH is stemmed from its theoretical assumption that traders are fully rational. It argues that investor reaction to the information news associated with the fundamental value causes a shift in the demand for financial assets; i.e. any change in the security prices is simply explained by the random arrival of fundamental news. The existence of the two heterogeneous agents namely; noise traders and arbitrageurs in capital markets (Black, 1986) make this argument difficult to sustain. To provide a clearer discussion about the reactions of the two different traders to information arrival, it is worth defining both types of investor. Noise trader is a financial term first introduced by Albert Kyle (1985) and Fisher Black (1986). Noise traders used to describe a stock investor who does not have any fundamental information about the traded security and lacks access to inside information. They make irrational investment decisions in financial markets hence the term irrational investors (De Long et al., 1990). On the other hand, arbitrageurs or as they are sometimes called (rational investors) are those who are always rational in making their investment choices. Rational investors are always updating their beliefs to reflect all publically available

Chapter Two: Literature Review

information. With these two definitions of noise traders (who traded on noise with no fundamental data) and arbitrageurs (who hold a rational belief) one would expect that the two types of traders would always possess an opposing opinion and would tend to trade against each other in financial markets. (More detailed information about the different roles of noise traders and arbitrageurs will be provided in section 2.3).

Investor sentiment plays an important role in determining asset prices. Black (1986) argues that investors frequently trade on noise, i.e. without possessing fundamental information, and often hold optimistic and pessimistic beliefs in determining assets' values (De Long et al., 1990). The effect of noise traders was first documented by De Long et al. (1990). They argued that the existence of two types of trader (noise traders and arbitrageurs) in the stock market has a great effect on assessing the security values. The unexpected change in noise investor sentiment may affect arbitrageurs' activities and prevent them from having opposing opinions to noise traders. The more noise traders become bullish or bearish toward a particular asset, the greater the price divergence of that asset from its fundamental value. This risk created by noise traders who cause changes in assets' prices away from their fundamental value is called "noise trader risk". The more a price deviates from its fundamental value, the riskier the security assets become.

In addition to the theoretical challenges, EMH has been empirically subjected to considerable criticism. A number of studies have provided empirical evidence that stock prices and returns can possibly be predicted from firms' past performances, market capitalisations and firm-specific financial ratios. For example, with a long-term horizon, Bondt and Thaler (1985) have proved that the predictability of the stock return from past performance of the firm is possible as returns exhibit a reversal effect in which shares with low performance in the past three to five years tend to have a higher return than shares with high performance over the same period. Conversely, on a short-term horizon, Jegadeesh and Titman (1993) showed that momentum may affect stock prices in such a way that the future behaviour of individual stock prices tends to follow the same directional movement of previous prices over the past six to twelve months. These studies (Bondt and Thaler, 1985; Jegadeesh and Titman, 1993) obtained results that contradicted the EMH and identified that stock price and return do not follow random walk behaviour. Another group of researchers has proved the predictability of stock returns from firm-specific financial ratios, such as price-to-

Chapter Two: Literature Review

earnings ratios (P/E), book-to-market value and cash flows. They found that firms with low market capitalisations and low P/E ratios yield a higher average return (Banz, 1981; Basu, 1977; Fama and French, 1988; Lakonishok et al., 1992; Chan et al., 1991). These empirical results, which contradict the theories, are called “market anomalies” and they are put forward as evidence of the inefficiency of financial markets. Stock market anomalies can be described as the irregularities of stock price behaviour of firms in financial markets (Levis, 1989). The existence of anomalies in the stock market (e.g., the momentum effect (Jegadeesh and Titman, 1993) challenges the foundation of EMH and suggests a new theoretical concept in modern financial literature, namely Behavioural Finance Theory. Over the last few years, financial economists have directed their attention to studying the effects of human psychological behaviour in influencing investors’ decisions.

2.2.5 From Traditional Finance Approach to Behavioural Finance Theory

The new paradigm of financial theory has come to play a complementary role in resolving the issues that traditional finance has failed to address. This new field in finance, called “Behavioural Finance” which is based on psychology, attempts to study the behaviour of people in financial markets. It is a new approach in the financial literature that has grown rapidly in recent years. It tends to address why people buy or sell financial assets through a psychological study of the characteristics of market participants that influence individual decision-making. It also focuses on how investors interpret and react to information when they make their investment decisions.

The existence of the different types of investors and the effect of their trading behaviour on price changes is considered one of the justifications for the assertion in this research that stock micro-blogging sentiments have predictive power in forecasting stock price movement in the stock market, which stems from behavioural finance literature. Behavioural Finance theory has challenged the EMH, which is based on the assumption that investors/traders are always rational. Behavioural finance theory, on the other hand, is based on the assumption that investors do not always hold rational beliefs (Baberis and Thaler, 2003). There are two types of traders in financial markets: the “irrational noise traders” (or so-called “liquidity traders” or “day traders”) and “rational traders” or “arbitrageurs”. Noise traders are

Chapter Two: Literature Review

those who trade on noise without possessing any fundamental information (Black, 1986; Glosten and Milgrom, 1985; Kyle, 1985). In contrast, rational traders or arbitrageurs are those who hold rational/Bayesian beliefs (DSSW 1990) (DeLong et al., 1990) since they always update and correct their beliefs to reflect any new information (Baberis and Thaler, 2003). They always trade against noise traders by taking advantage of the price difference by selling high and buying low. Behavioural Finance has come to explain various financial market anomalies. It also helps to explain the role of noise traders in describing investor behaviour and determining asset prices in the stock market.

2.3 The Role of Noise Traders in Capital Markets

Behavioural finance has provided evidence that noise investors' emotions play a major role in determining asset prices because of the arbitrary decisions they make regarding the sale and purchase of assets in capital markets (DSSW, 1990). It follows that investors' psychology, emotions, preferences and mistaken beliefs can affect the decisions of other investors on the market and may result in shifting the asset's value from its fundamental value.

The trading activities of noise traders in capital markets create an attractive profitable opportunity for arbitrageurs. They take advantage of the price difference and make profits by selling high (when noise traders push prices up) and buying low (when noise traders depress prices). Researchers argue that the psychological change in an investor's attitude in the market represented by his/her sentiments will have a great effect in altering the investor's decision and driving asset prices. Baker and Wurgler (2006) argue that, in practice, optimistic irrational investors, who trade more frequently and therefore add more liquidity, will bet on rising stocks more than pessimistic irrational investors will bet on falling stocks. This implies that, if investors are bullish (optimistic) about a particular stock, they are more likely to hold the stock for longer, which is a good signal to other investors to demand more of that particular stock, thus resulting in upward trends of stock prices. In contrast, if investors are bearish (pessimistic) they are likely to stay short and will tend to sell that stock, giving a bad signal to other investors not to buy that particular stock, thus driving down the asset's price. Hence, it can be clearly seen that the activities of noise traders will cause prices to fluctuate up or down around equilibrium, which will subsequently

Chapter Two: Literature Review

cause arbitrageurs to engage in trade by pushing prices back or forth around equilibrium, keeping the market efficient.

Noise traders' activities may create risk and subsequent limits to arbitrageurs. Whilst the arbitrageurs' actions are limited in the short run, these limitations will diminish in the long run when price deviations from fundamental/mean levels become sufficiently extreme to allow arbitrageurs to trade profitably against noise traders (McMillan, 2005). The extreme deviations may be a result of noise traders overreacting or under-reacting to good and bad news, causing price levels and risk to deviate far more drastically from expected levels than would have been actually required by the news. In spite of price deviations from fundamental values and limits to arbitrage, the profitable reversion trading strategies¹ of arbitrageurs may not be implemented immediately to avoid falling into a possible 'mispricing/misperception' trap (Shleifer, 2000). Arbitrageurs should have had correct market timing for their contrary trading strategies as the mispricing of noise traders becomes even more extreme as new fundamental information may unexpectedly arrive after an arbitrageur has taken his/her initial position. The changes in the noise trader's misperceptions affect asset prices. The misperception of the asset's risk causes noise traders to follow each other in selling (buying) risky assets just when other noise traders are buying (selling). The size of changes in the misperceptions of the asset's risk is responsible for the magnitude of the divergence in asset returns. The greater the misperception, the lower the expected return. Instead, an increase in misperception of the asset's risk will cause a rise in price uncertainty that deters risk-averse arbitrageurs from holding the risky assets. By moving away, the risk-averse arbitrageurs give noise traders advantages and the chance to gain a higher return from their trading.

2.3.1 Noise Traders and Stock Price Behaviour

Economists of behavioural finance suggest that noise traders have a significant influence in affecting stock price behaviour. The 'noise trader' model of DSSW (1990) shows that the irrational noise not only affects the equilibrium prices but also earns higher expected returns than those achieved by rational arbitrageurs. Noise

¹The reversion trading strategy is the strategy followed by arbitrageurs to bring the prices of securities back to the mean values. Arbitrageurs always trade against noise traders who tend to trade on noise and cause the price to move away (up/ down) from its mean values whereas the former (arbitrageurs) are said to follow a reversion strategy by bringing the security prices back to mean values.

Chapter Two: Literature Review

traders acting in concert can produce a systematic risk called ‘noise trader risk’ which deters arbitrageurs from trading against noise traders, resulting in the price deviating significantly from its fundamental value. This risk arises from the unpredictability of noise trader sentiment or opinion that limits the effectiveness of arbitrageurs in the capital market.

The DSSW (1990) model shows that there are four effects that might explain the influence of investor sentiment on stock prices. These effects are defined as follows:

- The first effect is the “hold more” effect which states that when noise traders on average hold more risky assets, they earn a larger share of returns to risk bearing; thus, their expected returns relative to those of arbitrageurs are increased.
- The second effect is the “price pressure” effect which states that as noise traders become more bullish, their demand for risky assets increases, thus driving up the price of those risky assets. Therefore, the return to risk bearing will be reduced, consequently reducing the discrepancy between their returns and those of arbitrageurs.
- The third effect is the “buy high-sell low” effect or the so-called “Friedman” effect which explains the misperceptions of noise traders. When noise traders hold stochastic beliefs that they have poor market timing, they tend to buy more risky assets just when other noise traders are doing so, meaning they are more likely to suffer a capital loss. This indicates that the effect of poor market timing on their returns will become greater as noise traders’ beliefs become more variable.
- The fourth effect is the “create space” effect, which serves as the central ‘noise traders’ model. This effect is also associated with the variability of noise traders’ beliefs, similar to the “Friedman” effect. It states that when the variability of noise traders’ beliefs increases, the price risk tends to increase. As arbitrageurs are risk-averse, they are more likely to bear this greater risk, which therefore reduces and limits their ability to trade against noise traders. DSSW (1990) suggests that two effects, the “hold more” and “create space” effects, tend to increase noise traders’ relative expected returns while the “buy

Chapter Two: Literature Review

high-sell low” and “price pressure” effects tend to lower noise traders’ relative expected returns.

Behavioural finance theory suggests that the existence of noise traders and arbitrageurs and their trading interactions have the power to affect trading activities and, hence, price formation in capital markets (DeLong et al., 1990). The trading activities of noise traders in capital markets will inevitably create an attractive, profitable trading opportunity for arbitrageurs. They take advantage of the price difference and make profits by selling high (when noise traders push prices up) and buying low (when noise traders depress prices). Researchers argue that the psychological change in investors’ attitudes in the market, represented by their sentiments, will have a great effect in altering their trading decisions and driving asset prices.

2.3.2 Investor Sentiment, Disagreement and Stock Market Relations

As discussed in previous sections, noise traders play a tremendous role in affecting stock prices in capital markets. In this section we discuss two specific issues that are deemed to be the central concern of this thesis. Does the bullishness of the message help to predict returns? Is disagreement among messages associated with more trades?

The first issue is whether the bullishness of messages (serving as a proxy for investor sentiment) predicts stock returns. Empirical studies have extensively addressed the role of investor sentiment in the formation of security prices in capital markets. Baker and Wurgler (2006) provide evidence from U.S. market sentiment that a broad wave of investor sentiment has a huge effect on assets whose valuations are highly subjective and more difficult to arbitrage. In examining the effect on security prices in the high/ low sentiment period, Karlsson et al. (2005) and Yuan (2008) find empirical evidence that sentiment traders participate and trade more heavily in high-sentiment periods than in low-sentiment periods. Barber and Odean (2008) show that since investors are reluctant to take short positions in the low-sentiment period, they have strongly demonstrated that their sentiments have a greater impact on prices during high-sentiment periods. Schmeling (2009) found a negative relationship between sentiments and return; i.e. when sentiment is high, the subsequent return tends to be lower, and vice versa. More importantly, it should be emphasised that

Chapter Two: Literature Review

most of the empirical findings have provided evidence of different market reactions to various levels of investor sentiment and that the effect of sentiments on stock prices can be asymmetric (Brown and Cliff, 2005; Gervais and Odean, 2001; Wang, 2001; Hong et al., 2000). These studies suggested that market performance during a period of growth (recession) induces an optimistic (pessimistic) attitude among investors about whether to speculate in the market. Specifically, DeBondt (1993) argues that increased bullishness might be expected after a market rise while increased bearishness might be expected after a market fall, thus confirming the hypothesis of “positive feedback traders”. Moreover, Verma and Verma (2007) have provided even more precise information on the existence of an asymmetric effect of the stock market on investor sentiments by emphasising the magnitude of effects of bullish (bearish) sentiments in different states of market innovation (growth and decline). Their findings reveal that variations in the stock market may have a stronger effect on bullish sentiment in a period of growth than similar effects on bearish sentiments in a period of decline.

The second issue is whether greater disagreement triggers increases in trading volumes of stocks. “It is the opinion differences that make a horse race” (Pfleiderer, 1984; Varian, 1985; Harris and Raviv, 1993) is the central concept leading this research study. Difference in opinion among traders has proven potential in determining asset prices in capital markets. While liquidity motives explain much of the variability in trading volume, the trading behaviours of noise traders, who trade on noise, and arbitrageurs, who hold rational beliefs, play a huge role in models of financial markets. Theoretical analysis has extensively investigated the extent of the effect of divergence of opinion on assets pricing. Explicit theoretical analysis of the relationship between investor disagreement and trading volume has long been undertaken by scholarly researchers. Copeland (1976) has provided a theoretical model that implies that the extent of disagreement among information recipients has an effect in determining trading volumes in the market. His model shows that the impact on trading volume in the form of a decrease (increase) depends on whether the information arrival is sequential (simultaneous). Later theoretical support is provided by Varian (1985) who documents a positive relationship between volume of trade and the degree of heterogeneous beliefs among investors in a simplified pure exchange setting. Similarly, Comiskey et al. (1987) find a significant positive relation between

Chapter Two: Literature Review

trading volume and the dispersion of opinion measured by the dispersions in analysts' forecasts of annual earnings.

A discussion of the broad body of literature on testing the validity of the Efficient Market Hypothesis and the Random Walk Theory has produced mixed results. In general, many early works tend to support the random walk model, while the majority of studies conducted recently showed contradictory results and have proved that stock price do not completely follow random walk behaviour and that some degree of prediction is possible. The emergence of behavioural finance theory and its explanation of the role played by noise traders in capital markets have successfully proved its complementary role with traditional finance theory.

2.4 Online Stock Forums

A growing body of empirical research has been undertaken to investigate the predictive power of online investing forums in predicting various financial market indicators; all of these papers have focused on message boards, financial news articles and, recently, on micro-blogging forums. The following section presents the related literature in each of these forums.

2.4.1 Internet Message Boards and Financial Markets

Internet message boards are one of the most popular investment forums providing an effective means for investors to communicate, disseminate and discover information (Delort et al., 2012). Recent studies have begun to explore the impact of stock message boards on financial markets and stock prices' behaviour (Wysocki, 1998; Tumarkin and Whitelaw, 2001; Antweiler and Frank, 2004a and b). The first paper to analytically investigate Internet postings was that of Wysocki (1998). He measures the correlation between the stock message volume and the next day's trading volume and stock returns in a sample of 50 most frequently discussed firms on Yahoo! Finance over a period of eight months in 1998. He shows in cross-section the difference between firms with high-volume message-posting activity characterised as high market valuation with high return and accounting performance, high volatility and trading volume, high price earning and market-to-book ratio, high analyst following and low institutional holding. Similarly, Das and Sisk (2005) analyse the

Chapter Two: Literature Review

high volume of stock message boards posted on a stock forum to determine how investors' opinions are linked and spread among tickers involved in that discussion forum. They found that high mean returns and lower return variance were highly associated with stock with high connectivity in the forum.

Tumarkin and Whitelaw (2001) were the first to investigate the directional link between stock message board activity, trading volumes, and stock market returns. However, their research study focused solely on quantitative data but used a different aspect of the stock message board contents, i.e. voluntary users' ratings. They concluded that users' ratings, from strong buy to strong sell recommendations, have failed to demonstrate whether these recommendations contain relevant information that might predict abnormal stock returns. Das and Chen (2007) used statistical and natural language processing techniques to classify the sentiments of 25,000 board messages on nine selected companies during the last quarter of 2000. They found evidence that sentiment is based on stock movements and can be used to predict future volume and volatility but cannot forecast future returns. The results of both studies, Tumarkin and Whitelaw (2001) and Das and Chen (2007), are consistent with the EMH, in that Internet message boards do not predict stock market returns.

One major criticism of the above-mentioned studies is that they rely too heavily on quantitative data from Internet message boards, such as message volume and users' ratings. Unlike previous works, the most complete study of Internet message boards is that of Antweiler and Frank (2004b), who focus on qualitative as well as quantitative data analysis of Internet messages posted on Yahoo! Finance and Raging Bull. They determined the correlation between activity on Internet message boards and stock volatility and trading volume. They show a minor correlation between message board posts and next-day price levels. They found that positive shocks to message board posting levels do predict negative stock returns on the next day. Another study, conducted by Koski et al. (2004), examined the stock message boards posted in Raging Bull and Yahoo! Finance of a large sample of NASDAQ stocks in 1999. They investigated the effect of noise trading on stock return volatility. Their study is in line with previous studies by Black (1986), DeLong et al. (1990), and Campbell and Kyle (1993) in testing the hypothesis that noise trading increases volatility. They found powerful evidence that trading increases volatility, which influences the volume of message board activity as a result.

Chapter Two: Literature Review

Although the above-mentioned studies explore the effect of the Internet message boards on financial markets, Jones (2006) sought to investigate the stock price behaviour pattern for companies in both periods, pre- and post-internet message boards. He highlighted that message boards “may be an observable form of a pre-existing information network or [...] they may have altered the information landscape in a way which has changed pricing behaviour” (p. 67). He examines the stock return behaviour of large numbers of firms listed in the S&P 100 both before and after their adoption of the Internet message board on Yahoo! Finance. He finds that the trading volume has increased significantly after the implementation of the Internet message board by the companies, while the daily stock returns were significantly lower in the period after message board adoption. As a result, this shows high variances in returns due to market risk associated with the new implementation of the internet message board. While the previous studies mentioned above carried little information content for future stock movements, Gu et al. (2006) have proposed an alternative approach to determine and combine information from millions of posts on stock message boards by using the weighted average recommendations of daily posts. They found that the evidence is both statistically and economically significant in supporting the hypothesis that the weighted average recommendation of stock message boards has predictive power with regard to future stock returns. While Gu et al. (2006) reveal message boards’ ability to generate abnormal returns, providing further evidence that new value-relevant information is generated via message boards, this result contradicts that of Dewally (2003) who finds no evidence that the recommendation exchange in online discussion groups has any informational value or that the postings have any impact on abnormal returns.

In spite of the large number of scholarly studies that have been attempted to address the predictive power of the stock message board over the financial market, this topic has always had a certain appeal for researchers. A number of research studies have been conducted recently to further investigate the impact of stock message boards on market prediction. These extended works have examined the impact of the stock message board from different aspects. For example, disaggregating message board information by type would allow more accurate and detailed analysis such as the study by Clarkson et al. (2006) who investigate the accuracy of rumours posted in the Internet Discussion Site (IDS) and market reaction

Chapter Two: Literature Review

to messages posted. Their findings suggest that rumoured earnings are considered a valuable source of information to predict future earnings. These results indicate that the market reacts keenly to investors' sentiment and discussion in the forum and to the volume of posts (Wysocki, 1998; Das and Chen 2001; Tumarkin and Whitelaw, 2001; Antweiler and Frank, 2004b). These results are consistent with those of Bettman et al. (2011) who have empirically investigated the impact of stock message board takeover rumours on equity market activity. By employing a computational linguistic method for a sample of 2,898 message board takeover rumours posted over a period of six years and utilising Intraday Trade Quote (TAQ) data, they found that the trading volume and the abnormal return are positively significant in relation to the message posts and rumours dissemination period. They have also examined the cross-sectional variation of the impact of stock message board takeover rumours in the US equity market by applying multivariate analysis; they found significant evidence of variation in relation to firm size, rumour rating, technology industry and prior media speculation effects.

Other researchers, such as Lerman (2010), have studied the impact of message boards by analysing the individual's attention to accounting information. He found that investors' discussions have increased mostly around accounting events such as earnings announcements (for example; quarterly report and annual report), which are in turn associated with a reduction in information asymmetry and better-informed investors. His results also revealed that more accounting-related discussions are highly associated with lower analyst coverage; higher analyst forecast dispersion and higher trading volume. Delort et al. (2012) have investigated the real impact of manipulation in online forums on the financial market. Their findings reveal that the message that has been manipulated due to ramping is significantly and positively related to stock market return, volatility and trading volume. They demonstrated that ramping is commonly associated with stocks with low market capitalisation, high turnover, low price level and high volatility of returns.

Overall, the aforementioned empirical findings highlight that message boards have proved to be effective channels for accelerating the dissemination of financial market information and facilitating communication amongst market participants (see, for example, Wysocki, 1998; Tumarkin and Whitelaw, 2001; Antweiler and Frank, 2004a and b; Das et al., 2005; Jones, 2006; Gu et al., 2006). Although the evidence

Chapter Two: Literature Review

generally suggests that message board activity has a certain impact on financial markets in some form, there are occasionally contradictory findings with regard to the predictive power of the stock message board in predicting stock price movement. These inconsistencies in the findings may be due to the massive amount of stock message boards examined (see, for example, Gu et al., 2006) or to inadequate methodological procedures, or both.

In general, the literature has demonstrated the effectiveness of the stock message board as a valuable investment forum and appears to have proved its great impact in predicting stock market variables such as abnormal return, trading volume and return volatility (Antweiler and Frank, 2002). Table 2.1 provides a summary of selected recent studies investigating the impact of stock message board in predicting financial markets.

Table 2.1: A summary of selected studies on Internet message boards

Study	Internet Message Board Activities		Financial Market features			Findings
	Message Volume	User's Ratings	Market Return	Trading Volumes	Return Volatility	
Wysocki (1998)	•		•	•		Strong positive correlation between stock message volume overnights and next day trading volume and stock returns.
Das and Chen (2001)	•		•		•	There is no evidence of the predictive capability of sentiment of stock message board in forecasting stock returns
Tumarkin and Whitelaw (2001)	•	•	•			Internet message board cannot predict stock market return.
Das and Sisk (2003)	•		•			High mean return and low return variance associated with stocks of high connectivity in the discussion forum.

Chapter Two: Literature Review

Dewally (2003)	•		•			No evidence found that the recommendations have informational value. Therefore, there is no evidence of the impact of posting on the cumulative abnormal return and the return appears not to be significantly above market performance.
Antweiler and Frank(2004)	•			•	•	There is minor correlation between internet message volume post and the next day price activity.
Koski et al. (2004)	•			•	•	There is evidence that trading noise increases volatility and affects message posts as a result.
Jones (2006)	•		•	•		The trading volume has increased significantly after the implementation of the internet message board by the companies while the daily stock returns were significantly lower in the period after message board adoption.
Gu et al. (2006)	•		•			There is statistical and economic evidence to support the hypothesis that the weighted average recommendation of the stock message board has predictive power with regard to future return of the stocks. However, the individual recommendations have no predictive power with regard to future stock return.

2.4.2 Financial News Articles and Financial Market

There is a vast amount of literature on the impact of financial news articles and investment stories on the stock market. Mitchel and Mulherin (1994) conducted

Chapter Two: Literature Review

an early study to investigate the impact of financial news articles on markets. This paper studies the relationship between the posting frequency of New York Times articles and Dow Jones announcements. They report a very weak relationship between media activity and trading volume and volatility. Gidofalvi (2001) presents an approach for investigating the relationship between financial news articles and short-term price movement. He extracted 5,000 news articles concerning 12 stocks. Each article is then classified as ‘up’, ‘down’ or ‘unchanged’ related to the movement of the stocks in a specified time interval twenty minutes before and twenty minutes after the articles are released to the public. He shows a strong correlation between news articles and stock price movement during the time interval specified. A similar pattern of predicting price behaviour in specified time intervals is adopted by Schumaker and Chen (2009). They investigate the prediction of the discrete price value using textual analysis of 9,211 financial news articles. They construct a model to estimate the stock price movement twenty minutes after the news articles are publicly released and validate the importance of news for the performance of a stock. Likewise, Lavrenko et al. (2000) build a language model to predict the future behaviour of the stock by checking the language model of the news that occurred in previous hours.

Tetlock (2007) studies the news content of the Wall Street Journal’s (WSJ’s) “Abreast of the Market” to analyse how news sentiment affects the daily stock market activity and whether the WSJ’s content can affect stock market return. He measures the sentiments found in the WSJ by creating a simple measure of pessimism to determine inter-temporal relations with the stock market. He found that media content has a predictive ability to predict movement in the stock price and return and trading volume. He argues that high trading volume will be associated with both unusual increases and decreases in media pessimism. This study was extended by Tetlock et al. (2008) to analyse the effect of the negative words in all Wall Street Journal (WSJ) and *Dow Jones News Service (DJNS)* stories about individual S&P 500 firms from 1980 to 2004. Their main results are that the negative words in financial news and stories predict low firm earnings and downward movements in the firm’s stock price. They argue that the predictive power of the negative words is greater in the stories that focus on the fundamental aspects of the firms.

Chapter Two: Literature Review

Seeking Alpha (www.seekingalpha.com) is the most popular investment website forum, providing a valuable source of financial news articles. Fotak (2007) has addressed the effect of Seeking Alpha long-term and short-term stock recommendations on related prices and investors' volume traded. He found that when the recommendations on a particular stock were given by a professional analyst with a degree in a subject such as economics and finance, the general price and volume of trade were more substantial than when the recommendation was provided by non-professional analysts. Most recently, Chen et al. (2013) investigated whether investors' opinions shared on social media platforms have any power in predicting financial market variables such as earnings surprises. They applied textual analysis techniques to the WSJ financial news articles released in the most popular investment website forum, Seeking Alpha (www.seekingalpha.com). They argued that the views that are mostly expressed in these articles are strongly associated with subsequent and contemporaneous stock returns. They found that the articles' contents posted on Seeking Alpha are value-relevant in predicting earnings surprise.

Beyond the financial news articles published in the Wall Street Journal, a number of research papers have studied the impact of financial text released on other multimedia platforms. For example, Davis et al. (2012) examine sentiments of earnings press releases on PR Newswire. A sample of 23,000 quarterly earnings press releases publicly available on PR Newswire for the period 1998-2003 have been collected to measure net optimistic language. They employ textual analysis to classify the language in the press releases as optimistic or pessimistic. They show that the net optimism of the quarterly press releases has predictive power regarding firm performance in future quarters. Similarly, Loughran and MacDonald (2011) make use of textual analysis of financial text to measure the tone and sentiment of corporate 10-K reports during the period 1994-2008 by employing the Harvard Dictionary. Their findings reveal that three quarters of the words classified as negative in the Harvard Dictionary do not appear to have the same negative tone in the financial context, while there is evidence that some word lists are correlated with stock market reaction around 10-K released data, unpredicted earnings, trading volume and return volatility².

²10-K report is a form of annual report required by the U.S. Security Exchange Commission (SEC), which provides a comprehensive summary report about company's financial performance. It contains much more detailed information than the normal annual report. It includes data such as company history, organizational structure, equity, earning per share, etc.

Chapter Two: Literature Review

While numerous studies have been conducted to address the effect of stock message boards, financial news articles and investment stories on predicting financial markets, these studies generally find significant market reaction to stock message board posts as well as financial news. However, the recent advance of investment micro-blogging forums as a medium of communication for investors, financiers and market analysts is attracting most researchers' attention nowadays. This is a fruitful attempt by scholars to address the impact of such financial micro-blogging forums as little is known about this topic.

2.4.3 Micro-blogging Forums and Financial Market

Recently, with the pragmatic innovation of online investment forums around the world, platforms such as StockTwits and TweetTrader have become widely used online discussion forums among investors and traders. A small number of empirical papers have analysed the impact of the information content of micro-blogging on predicting the stock market. A number of studies have been conducted to investigate the impact of Twitter messages' sentiments at macro- and micro-economic levels.

In predicting macro-economic indicators, O'Connor et al. (2010) analyse the relationship between sentiments of Twitter messages and public opinion to investigate whether sentiments can help to predict a consumer index. They analyse a survey on consumer confidence and public opinion and show a strong correlation between Twitter sentiment scores and public opinion over time, signifying that automatic sentiments on Twitter could help in observing public opinion about a particular topic. They have found evidence that Twitter message sentiments are a leading indicator of a consumer index (such as the U.S. elections and consumer confidence). Meanwhile, other studies have been conducted to explore the effect of Twitter sentiments on a micro-economic market indicator. For example, Bollen et al. (2011) produced the first paper to analyse the predictive power of Twitter sentiments at a micro-economic level in forecasting the performance of Dow Jones Industrial Average Index (DJIA). They measure the average mood of public users of Twitter and associate overall mood with future DJIA return. They show that some moods such as happiness and calm have some predictive power regarding stock market volatility and market return. Zhang et al. (2011) show how the emotional effect of public mood on Twitter is reflected in

Chapter Two: Literature Review

movements of stock market indicators such as the S&P 500 and NASDAQ. Both studies have provided empirical evidence of the predictive power of public tweet sentiments in predicting market indices. Research by Baik et al. (2015) takes a slightly different approach by focusing on the Twitter user location (local and nonlocal Twitter users) rather than on the message volume or message sentiments. Their study reports that local Twitter users significantly predicts future stock returns while they fail to confirm any predicted ability of returns by nonlocal Twitter users.

A few studies specifically focus on the specific domain of stock micro-blogs and investigate their ability in predicting stock market indicators. Sprenger et al. (2014) investigate the relationship between market prices of publicly traded companies and the StockTwits sentiments. They show that the sentiment of StockTwits (i.e. bullishness) is significantly associated with abnormal stock returns and message volume while that sentiment has power in predicting the next-day trading volume. Oh and Sheng (2011) study the predictive power of stock micro-blogging sentiment in forecasting stock price directional movement. They find that the real-time features of stock micro-blogging have predictive power regarding future stock price movement. Rao and Srivastava (2014) studies the correlation between Twitter sentiments and stock prices and they found a significant correlation between stock prices and twitter sentiments while confirming the short-term effect of Twitter discussions on stock prices and indices movement. Ranco et al. (2015) used StockTwits of 30 DJIA companies employing time series data for each of these companies. They conducted an “event study” techniques which a replicate of similar study of Sprenger et al., 2014 and have come to similar conclusion that both twitter volume and sentiments contained value relevant information in predicting stock returns.³

Having discussed various online investment forums that enable investors and financial professionals to share and exchange investment ideas and opinions in order to make better-informed investment decisions, it can be argued that the information

³Although Ranco et al., (2015) used relatively similar data of StockTwits of 30 DJIA companies, our study is different in several ways. First, the data used in this study is structured as a panel data with company fixed effect (controlling for company specific characteristics such as size, etc.) unlike Ranco et al., (2015) who employ time series data for each company of DJIA. Second, our study adopts a more robust evaluation of the usefulness of sentiment measures of microblogging data for predicting stock returns (i.e., employing QR techniques that examine the relationship of sentiments and returns over spectrums of conditional quantiles of returns) (Refer to chapter 7 for more details). Ranco et al., (2015), by contrast, conducted an “event study” techniques which a replicate of similar study of Sprenger et al., 2014. Third, we provide an in depth analysis of the effect of sentiments on returns by investigating the asymmetrical behavioral impact of investor sentiment on financial variables (i.e., returns, trading volume and volatility) by distinguishing between the bullish and bearish sentiments.

Chapter Two: Literature Review

posted on online stock investment forums is a highly valuable source of data for making accurate stock market predictions. It is therefore vital to adopt certain technical methods to extract sentiments from those texts posted on online stock forums. Data mining techniques for sentiment detection of opinion from online text is a field of research that has received significant attention from many researchers in recent years. Financial economists and stock market analysts have greatly realised the importance of text mining for the purpose of stock market prediction. The next section presents the various tools and techniques used for text mining and discuss the different machine learning algorithms for sentiment extraction.

2.5 Text Mining

There is a growing body of theoretical and empirical research addressing text mining and its relative importance for extracting meaning from texts for various purposes. The following subsections define sentiment analysis and discuss various methods and approaches through which sentiments and/or opinions on a topic or an issue can be extracted from a text, sentence and phrase or overall context of a document.

2.5.1 Text Mining Definition

Text Mining (TM) is a field of study that has evolved to address the potential issues of managing and handling massive amounts of information to extract meaning and knowledge from text. In terms of handling large quantities of information, text mining can be defined as “the process of extracting useful information from textual databases through the application of computer based methods and techniques” (Fan et al., 2006). Meanwhile, in terms of extracting knowledge and meaning from textual databases, text mining can also be defined as “the non trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in unstructured data” (Karanikas and Theodoulidis, 2002). Sentiment analysis builds intensely upon text mining to extract useful information from text. Sentiment analysis (also known as opinion mining) refers to the extraction of the implicit meaning that was previously unknown from data or text (Witten et al., 2011). This research study attempts to use the terms ‘sentiment analysis’ and ‘text mining’ terms more or less interchangeably.

Chapter Two: Literature Review

The primary aim of sentiment analysis is to identify the opinions, attitudes or thoughts of a speaker or a writer with regard to certain topics or overall context of a document.

Text mining performs nine different tasks for dealing with and handling rich sources of information and extracting knowledge from textual databases. These tasks or functions range from extracting useful information to visualisation and are categorised as information extraction, text-based navigation, search and retrieval, clustering, categorisation, summarisation, trends analysis, associations, and visualisations (Fan et al., 2006; Singh et al., 2007; Gupta and Lehal, 2009). One of the most widely used functions of textual analysis is to predict trends and future patterns based on time-dependent data while associating these patterns with other patterns extracted from the data under analysis. When text mining is used to perform such functions, it is defined as a subfield of data mining techniques. Therefore, it inherently requires techniques from other fields of Information Retrieval, Data Mining and Computational Linguistics (Bolasco et al., 2005), as shown in Figure 2.1.

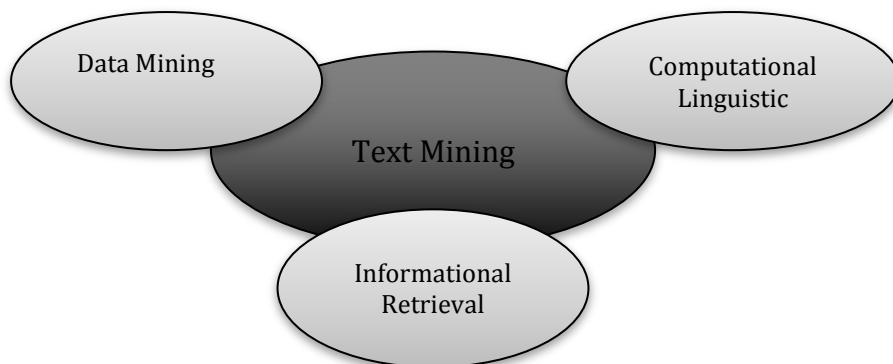


Figure 2.1: Text mining as an interdisciplinary field

2.5.2 Text Mining Tasks

Text mining performs various tasks to extract meaning from textual data. For example, some tasks aim to detect sentiment polarity of text, some are concerned with sentiment strength, while others perform feature selection methods to identify the most relevant features from datasets. These are presented in the following sections.

- **Sentiment Polarity**

The essential task in sentiment analysis is to identify the polarity of given texts in documents, sentences or phrases (Wilson, 2005). Sentiment polarity refers to the

Chapter Two: Literature Review

classifying of texts or emotions expressed in texts as either subjective/objective, positive/negative/neutral or any other polarity level such as emotion dimensions. This method of determining the polarity of the text as positive, negative or neutral has been widely used in studies of sentiment-related issues. Previous studies (Pang et al., 2002; Turney, 2002) have been carried out in that area and have applied different approaches for identifying the polarity of movie reviews and product reviews (e.g. automobiles, banks, movies and travel destinations). A study by Liu et al. (2005) performed sentiment analysis to detect the polarity (negative and positive) of customer opinions on competing products (digital cameras) on the web. A similar study by Gamon et al. (2005) applied a sentiment approach to analyse a consumer's car review to detect the polarity of expressed consumer opinions in the text.

Identifying the subjectivity and objectivity of a given text is another approach to sentiment polarity. The basic task of this method is to classify the text or document into one of two categories: objective or subjective. Pang and Lee (2008) argue: "the problem of distinguishing subjective versus objective instances has often proved to be more difficult than subsequent polarity classification" (2008, p. 977). Determining the subjectivity of a document may be misleading as an objective document may contain a subjective word or phrase. Moreover, Su and Markert (2008) noted that, when determining the subjectivity of a word sense in a text, the subjectivity is not tailored towards a specific word sense and the result with different resources varies substantially depending on the context definitions of subjectivity used when mining texts.

The use of sentiment and textual analysis to automatically measure emotions and extract attitudes, opinions and sentiments has been growing rapidly in the last few years (Wilson et al., 2006). For example, Strapparava and Mihalcea (2008) performed automatic analysis of emotions in texts of news headlines extracted from news websites. They used six basic emotional states - anger, disgust, fear, joy, sadness and surprise - while proposing and evaluating several knowledge-based methods for automatic extraction of these emotions in text. Another research study, by Dodds and Danforth (2010), sought to determine the average level of sentiment as well as the overall trends in the level of happiness for a diverse set of large-scale texts: song titles and lyrics, weblogs, and State of the Union addresses. Kramer (2010) adopted a similar approach of algorithmic extraction of a user's emotional state by using

Chapter Two: Literature Review

Facebook status updates to track changes in mood over the year and to measure “the overall emotional health of the nation”. Another study, by Bollen et al. (2010), performed sentiment analysis of public tweets by extracting six dimensions of mood (tension, depression, anger, vigour, fatigue and confusion). Table 2.2 provides a summary of different research studies that adopt sentiment polarity in different contexts by using variety of opinions and or emotional measurements.

Table 2.2: A Summary of selected research studies on sentiment polarity.

Study	Context	Measurement	Polarity/ Classification
Kramer (2010)	Facebook	Emotional health of the nation	Positive /negative
Dodds and Danforth (2010)	Blogs, Song lyrics, Song titles, State of the Union addresses	Level of happiness	(love, hate, pain, fear, life, truth, death,.....)
Bollen et al. (2010)	Twitter	Mood dimensions	(tension, depression, anger, vigour, fatigue, confusion)
Strapparava and Mihalcea (2008)	News Website/ news headlines	Emotional states	(anger, disgust, fear, joy, sadness and surprise)
Pang and Lee (2008)	News Articles/blogs	Text opinion	Subjectivity/objectivity
Gamon et al. (2005)	Customer car reviews	Customer opinions	Positive/negative/neutral
Liu et al. (2005)	Websites (online forums, discussion groups)	Customer opinion	Positive/negative
Turney, 2002	Product reviews (e.g. automobiles, banks, movies, and travel destinations)	Customer opinion	Excellent/poor

- **Sentiment Strength**

Despite the wide acceptance of algorithmic sentiments in detecting polarity of the text, this application would not be sufficient in other applications where the texts often contain a mix of positive and negative sentiments (Thelwall et al., 2010). Therefore, an expansion of the basic task for detecting the polarity of a given text as

Chapter Two: Literature Review

positive, negative or neutral to determine the polarity strength in the text or a document would be more appropriate in some applications to detect both polarity and strength simultaneously. Sentiment strength algorithms attempt to use a numerical rating scale to indicate the strength of any sentiment detected.

Several studies have been conducted to determine sentiment strength in a given text or document. For example, Pang and Lee (2005) applied numerical ratings (to 3- or 4-star scales) to detect sentiment strength of rating-inference problems rather than simply determining whether a review is ‘thumbs up’ or not, while Snyder and Barzilay (2007), in a study of restaurant reviews, produced a set of numerical scores for various aspects of a given restaurant (food, service and atmosphere). Other research work by, Wilson et al. (2006), used a Support Vector machine learning algorithm (see Section 2.5.4) to classify the intensity of opinion in order to identify the weak and strong opinion clauses in text. In addition to numerical ratings, a new approach to sentiment strength was adopted recently by Thelwall et al. (2010). They based their approach on the fact that one can differentiate between the strong and soft emotions in the text. For instance, ‘love’ may be regarded as a stronger positive emotion than ‘like’.

2.5.3 Feature Selection

Feature selection has come to be used to refer to the process of determining the subset of features that should be selected from the data and presented for prediction and classification algorithms (Jones and Smith, 1991). Feature selection is found to be an essential pre-processing step in the text mining process. Feature can be defined as a characteristic, an attribute or an aspect of something whereas in this context of study a feature might also refer to a prominent (relevant) term in a given text. The primary purpose of feature selection is to eliminate the effect of irrelevant features from the database while retaining the features relevant to the classification problem (de Souza et al., 2006). By removing the features that have no discriminatory power (John et al., 1994), one can perform the classification in a cost-effective and time-efficient manner, often leading to more accurate classification results (Guyon and Elisseeff, 2003; Yang and Olafsson, 2006). Feature selection has been effective in reducing the dimensionality of the data as well as enhancing comprehensibility and generalisability (Zheng and Zhang, 2008).

Chapter Two: Literature Review

Feature selection algorithms are based on the belief that all forms of data are composed of two types of set features: relevant and irrelevant. Relevant features are those that hold valuable information about the classification problem. In contrast, irrelevant features contain no useful information about the classification problem; therefore, excluding them improves the classification performance, potentially leading to more accurate classification results. With so many features extracted from the data sets, researchers have realised (Yang and Pederson, 1997; Xing et al., 2001) that it is normal for some of these features not to be informatively important with respect to a given class or category, as if they are irrelevant or redundant.

Feature selection has attracted the attention of many researchers in several fields such as statistics (Miller, 2002), machine learning (Liu et al., 2002; Robnik-Sikonja and Kononenko, 2003), data mining (Kim et al., 2000; Dash et al., 2002) and stock market prediction (Huang and Tsai, 2009; Ni et al., 2011; Zhang et al., 2014). It has proved successful in several data mining tasks such as classification (Dash and Liu, 1997) and clustering (Dash et al., 2002; Xing and Karp, 2001). Several studies have investigated feature selection in machine learning. For instance, Liu et al. (2002) considered the issue of active feature selection through the application of the feature selection algorithm Relief. Robnik-Sikonja and Kononenko (2003) theoretically and empirically investigated various features, parameters and different uses of Relief algorithms. They explained critically how and why they work through theoretical and practical analysis of their parameters. Zhang et al., (2004) employed ReliefF algorithms for feature selection. They found that the Naive Bayes classifier based on ReliefF algorithms is sufficiently robust and efficient to preselect active galactic nuclei (AGN) candidates. Feature selection is commonly used in the area of stock prediction. A considerable amount of literature has applied different feature selection methods to predict stock price movements. For example, Xue et al., (2007) adopted the classification complexity of SVM as a feature selection criterion to predict the Shanghai Stock Exchange Composite Index (SSECI). Huang et al., (2008) employed a wrapper approach to select the optimal feature subset and apply various classification algorithms to predict the trend in the Taiwan and Korea stock markets. Lee (2009) proposed a prediction model based on a hybrid feature selection method and SVM to predict the trend of the stock market.

Chapter Two: Literature Review

Two different methods are commonly used to perform feature selection; they can be broadly characterised as the filter method and the wrapper method. The difference between the two methods can be explained in two ways. First, the methods differ in terms of the way in which the relevancy of the features is evaluated. Second, they may vary in terms of the exact time when the feature selection process occurs. The feature selection may occur inherently with the classification rule, as with the wrapper method, or it may occur before the standard classification rule performed to a subset of features, as with the filter method (Sima and Dougherty, 2008). The filter approach is usually based on general characteristics of the data sets (Liu and Yu, 2005) and statistical criteria, which scores the relevancy of the features. The features are then ranked in accordance with their relevancy where the most relevant features will be at the top of the ranking lists and relevancy will decline towards the bottom of the lists (Huang and Chow, 2007). The selected features under the filter approach are independent of the classifier algorithms (Yu and Liu, 2004). The wrapper method uses the learning algorithms to determine the relevant features. The features that are scored relevant in a given classifier algorithm may not be relevant in other learning algorithms. Further explanation of filter and wrapper methods is provided in the following sections.

(A) Filter Method

The filter approach (Dash et al., 2002; Yu and Liu, 2004) is the most commonly used method for feature selection tasks. It is based on filtering out the irrelevant features and returning the most relevant ones (Kohavi and John, 1997). It employs statistical measures to score the relevancy between the subset features and the class label (Liu and Yu, 2005). These measures vary in complexity from simple correlation measures, such as Pearson's Correlation Coefficient (Pearson, 1901), to complex correlation measures such as the Relief algorithm (Kira and Rendell, 1992). These correlations measure the informational strength and the predictive power of the relationships between the subset features and the class (Zheng and Zhange, 2008). Then, a relevancy score calculated by a statistical measure (either high or low) is assigned to each feature according to its significance (Liu and Yu, 2005). The features are then systematically ranked in order, with the top features with the highest relevance scores being selected and used by the classifiers and the low-scoring

Chapter Two: Literature Review

features being discarded (Sayes et al., 2007). The filter measures have some general characteristics, as suggested by Zheng and Zhang (2008):

- The selected features under the filter approach are independent of the classifier algorithms.
- The filter methods consider each feature in the data set independently.
- The process of identifying relevant features is relatively fast.
- Redundant or irrelevant features may be included, especially those measuring univariate relationships.
- Interactions among features are not considered. Some features have a strong discriminatory power when considered in combination with other features in a group while appearing weak as individual features.

(B) Wrapper Method

In contrast to filters, wrapper methods (Kim et al., 2000; Kohavi and John, 1997) are classifier-dependent (Lee, 2009). The relevancy of the subset features are evaluated and scored based on the classification accuracy of a classifier algorithm. In the wrapper method, the induction algorithm is used as a black box (Kohavi and John, 1997) to conduct the feature selection and search for good subset features.⁴ As a result, the selected features are strictly fitted to the classifier algorithm used. This means that the selected features that are found to be best in one algorithm may not perform well if used with other classifiers. The wrapper method is said to be an iterative process as the classifier is run in a repetitive manner. In each repetition, the classifier is run using a different subset of the original features on the dataset. Then, performance evaluation methods (e.g. cross-validation) are used to evaluate the accuracy of each subset (John et al., 1994). As with the filter approach, the relevant feature of the highest accuracy score will be used as the input to the classifier algorithm.

There are two approaches to subset selection involved in the wrapper method: forward selection and backward selection. According to Kohavi and John (1997), the term ‘forward selection’ refers to a search that begins with no features and successively adds more features that are deemed relevant by the classifier, whereas the backward selection refers to a search that begins with all the features and deletes

⁴The induction Algorithm is a set of formal rules extracted from a set of observations that may represent a model of the data.

Chapter Two: Literature Review

features considered redundant by the classifiers. The forward selection approach is the most widely used approach in the wrapper method as it considered much faster and less expensive than the backward selection approach (Kohavi and John, 1997).

As with filter methods, wrapper methods have some general characteristics. The characteristics of wrapper methods are listed below:

- The wrapper approach can be applied to any type of machine learning.
- The wrapper method is a computationally expensive and time-intensive process because of the repetitive involvement of the classifier algorithm for every subset of features in the dataset.
- High classification accuracy of subset features is produced by the wrapper.
- The relevant features are strictly tight to a single classifier. Features deemed relevant in one classifier may be irrelevant in other classifier algorithms.

In general, different classifier algorithms can be used with the wrapper method to perform feature selection. Three classifier algorithms are most commonly used to perform feature selection with the wrapper method in different areas of research: Naïve Bayes classifiers, Decision Tree classifiers, and Support Vector Machines. The wrapper method has proved successful in three fields of research, which have received great attention by researchers in the field: web mining (Stein et al., 2005), financial analysis (Ni et al., 2011; Huang et al., 2008), and bioinformatics (Li et al., 2004).

As it will be explained in Chapter 4, in this research study, all three aforementioned classifier algorithms will be used to perform feature selection for the purpose of predicting stock market trends and movements. The next section provides an overview of the three algorithms of interest (Naïve Bayes, Decision Tree, and Support Vector Machine).

2.5.4 Text Mining Techniques

Different methods have been developed using an algorithm to automatically extract meaning from text or a given document. These methods vary according to their task in detecting sentiment. Many useful approaches to data mining techniques are used frequently for textual analysis and sentiment detections (e.g. Decision Tree

Chapter Two: Literature Review

Classifier (DT), Naive Bayes Algorithm (NB) and Support Vector Machine (SVM)). A brief description of each of these approaches is provided below:

- **Naive Bayes Algorithm**

Naive Bayes Classifier is a data mining technique that has proved effective in many practical applications, especially in the text classification field (Rish, 2001). It is a joint probability distribution (Pearl, 1988) based on the Bayesian theorem. Naive Bayes is a classifier that belongs to the Bayesian Network (BN) family. A Bayesian Network (BN) describes the joint probability distribution (a method of assigning probabilities to every possible outcome over a set of variables, $X_1 \dots X_N$) by exploiting conditional independence relationships represented by a Directed Acyclic Graph (DAG) (Pearl, 1988). Please see Figure 2.2a for an example of BN with five nodes (X_1, X_2, X_3, X_4 and X_5). Each node in the DAG is characterised by a state, which can change depending on the state of other nodes and information about those states propagated through the DAG. This kind of inference facilitates the ability to ask 'what if?' questions of the data by entering evidence (changing a state or confronting the DAG with new data) into the network, applying inference and inspecting the posterior distribution (which represents the distributions of the variables given the observed evidence). For example, one might ask: 'What is the probability of seeing a strong growth in the stock market if the terms "bullish" and "confident" are commonly seen in tweets'?

There are numerous ways to infer both network structure and parameters from the data. Search-and-score methods to infer BNs from data have frequently been used. These methods involve performing a search through the space of possible networks and scoring each structure. A variety of search strategies can be used. BNs are capable of performing many data analysis tasks including feature selection and classification (performed by treating one node as a class node (C) (the category where each feature belongs)) and allowing the structure learning to select relevant features (Friedman et al., 1997) (See Figure 2.2b).

Chapter Two: Literature Review

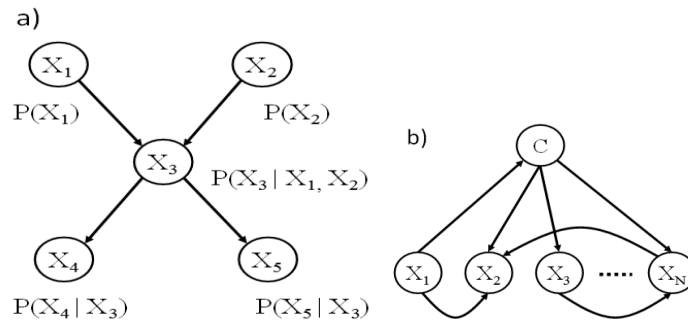


Figure 2.2: a) A Simple Graphical representation of a Bayesian Network with five nodes and b) a Bayesian Classifier where C denotes the class node

However, the naive assumption assumes conditional independence in which each feature in the database holds a mutually exclusive state of being independent of all other features. It considers each attribute (feature) separately and independently when classifying new incoming instances. Although the occurrence of each attribute is independent of the others, all attributes contained in the dataset are equally important.

Naive Bayes Classifiers possess attractive properties that have led many researchers to adopt them in several data mining tasks. Numerous researchers have used Naive Bayes with the wrapper method to select relevant feature subsets, mainly from bioinformatics data, particularly gene expression data. For example, Vinciotti et al. (2006) used the Naive Bayes Classifier technique with wrapper to select subsets of the most relevant genes from two genes databases (Prostate cancer database and B-cell lymphoma cancer database) to help diagnose cancer. Their result reveals that the wrapper approach to feature selection was able to extract a small number of genes from both genes databases that were deemed the most relevant. They found that this small number of selected genes reported a high accuracy level and enhanced biologists' understanding of correlations between the selected genes and cancer. Abraham et al. (2007) performed a similar series of experiments (using Naive Bayes with wrapper method) to show how the wrapper was able to extract relevant genes from 17 well-known bioinformatics databases. Their experimental results showed that the small number of relevant genes selected by the wrapper method by Naive Bayes led to higher classification accuracy than would have been the case had all the genes been selected.

Chapter Two: Literature Review

The Naive Bayesian classification method is the most widely used algorithm in the area of online financial text classification. Das and Chen (2007) employed the Naive Bayes Classifier to extract relevant investor sentiments from stock message boards. Their method is based on the word count of positive- and negative-association words where each word in a message is checked against the lexicon and assigned a value (-1, 0, +1) based on the default value (sell, null, buy) in the lexicon. Antweiler and Frank (2004b) and Sprenger et al. (2014) used the Naive Bayesian Classifier to classify messages automatically into three distinct classes (sell, buy or hold) based on the conditional probability of the words occurring in a particular class.

- **Decision Tree Classifier**

A Decision Tree (DT) is a technique that has been widely used for classification and prediction problems in data mining (Mitchell, 1997; Polat and Guneş, 2007). It is a tree-like graph made up of nodes, branches and leaves (Freitas, 2002). The purpose of the DT is to clarify the relationships embodied in the data by subdividing instance variables within the dataset (Chien et al., 2007). The separation of instant variables depends on the values assigned to one or more attributes. In the separation process, certain criteria are used to determine the relevance of the features and/or attributes with respect to the target variable (Chang, 2007). In the decision tree, each node represents the attribute name, the branch indicates the attribute value and, finally, the leaf indicates the predicted class. The separation process continues until the splitting of data reaches some predetermined level where the classifier results in a graphical representation in the form of a hierarchical tree structure (White and Sutcliffe, 2006). Figure 2.3 shows how the decision tree performs the classification function by building a decision node (attributes; e.g., A_1 and A_2) from a set of training data and further partitioning the nodes into sub-nodes (branches; e.g., $V_{1.1}$, $V_{1.2}$, $V_{1.3}$,) that represent the value of the attribute. The value will be assigned to the attribute based on the information gain criteria (IG). The partition process ends with the bottom leaves where the predicted class (e.g., class C_1 and class C_2) is assigned to the object's attributes. The selection of an attribute at each decision class will be the one with highest information gain. The decision rule extracted from the decision tree graph is based on "if- then" rules. For example, if the attribute A_1 has a value of $V_{1.1}$ then, it will be assign to the decision class C_1 whereas, if A_1 has a value of $V_{1.3}$ then, it

Chapter Two: Literature Review

will be assign to the decision class C_2 . There are some cases where an attribute is connected to another decision node that needs to be further split into sub-nodes as indicated by the attribute value $V_{1,2}$ where it is connected to attribute A_2 that is split further into two branches that takes $V_{2,1}$ and $V_{2,2}$ that are connected to a decision class C_2 and C_1 respectively.

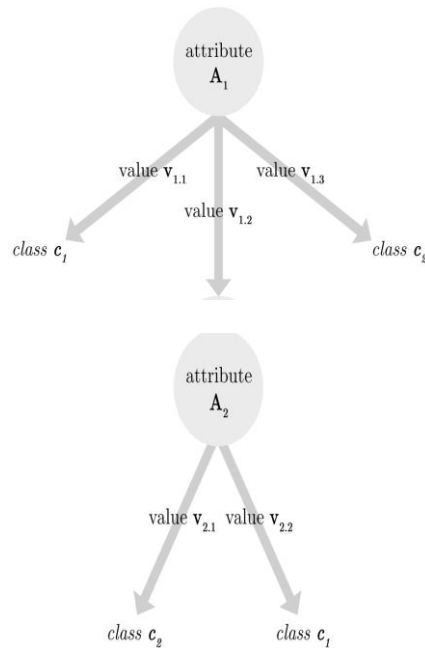


Figure 2.3: A Decision Tree Structure

In Figure 2.3, the relevant features are clearly identified with their respective relationships with other features in the datasets. The easy visualisation of the relationships among relevant features makes DT an attractive classifier, which has been used by several researchers in various fields such as bioinformatics, the web and the stock market. In the bioinformatics field, a study by Li et al. (2004) developed a decision-tree approach to perform multiple gene mining tasks through the efficient use of wrapper methods. They applied a wrapper approach to select the most significant and relevant genes that help predict cancer. By analysing two publicly available databases (colon data and leukaemia data), they were able to identify 20 highly significant colon cancer genes and 23 highly relevant leukaemia genes. These genes were found to generate a very high accuracy level compared to the accuracy generated if all genes are selected. With regard to web mining, Stein et al. (2005) adopted the wrapper method to select relevant features that may help identify the characteristics of hacker attacks on the Web. They made use of the well-known C4.5

Chapter Two: Literature Review

decision tree to perform the wrapper method to evaluate the relevancy of subset features. They found that a small number of relevant features were able to determine the main characteristics associated with web attacks. They showed that these relevant features selected by the decision tree resulted in a higher accuracy rate than would have been achieved using all features.

Decision tree algorithms have been used extensively in the area of stock market prediction. Wu et al. (2006) combine the filter rule and decision tree technique to develop a trading mechanism to screen effective buying points with higher average returns. Lai et al. (2009) constructed an investor decision support system based on a financial time-series forecasting model by integrating a data clustering technique, a fuzzy decision tree (FDT) and a genetic algorithm (GA) using data from the Taiwan Stock Exchange. Later, in 2011, Chang, Fan and Lin published a paper in which they made use of a case-based fuzzy decision tree model in an attempt to extract fuzzy decision rules that could be used as a basis for future time-series predictions. Their study aimed to provide investors with a decision support mechanism that would help them make better future decisions based on current market conditions. Chang (2011) conducted a comparative study of stock price prediction models using three prediction models: artificial neural networks (ANN), decision trees and a hybrid model of ANN and decision trees (hybrid model). His findings supported the ANN model and suggested it as a promising model in stock price prediction in the volatile post-crisis stock market.

- **Support Vector Machine**

The Support Vector Machine (SVM) is one of the most effective classification techniques originally introduced in the 1960s by Vapnik (1963) and Vapnik and Chervonenkis (1964). It has proved effective in both linear and non-linear classification problems. It is basically derived from Statistical Learning Theory (Cortes and Vapnik, 1995; Vapnik, 2000). The primary purpose of SVM is to maximise the hyper-plane that separates the data instances of two classes (Stitson et al., 1996; Barakat and Bradley, 2007) as accurately as possible, as shown in Figure 2.4

Chapter Two: Literature Review

SVMs are powerful classification techniques commonly used for linear as well as non-linear separable data. In linearly separable classifications, the SVM model represents the instance in a given class, as points in space, which are tolerably mapped and optimal hyper-planes, are found to widely separate the data instances of the two classes in the same space. However, in non-linear classifications, the SVM makes use of kernel functions (Chen and Hsieh, 2006) to transfer the data instances from input space into high-dimensional feature space so that the data are linearly separable (Saunders et al., 1998). In addition to classification problems, SVMs work effectively in regression and time-series prediction applications (Smola and Scholkope, 2004).

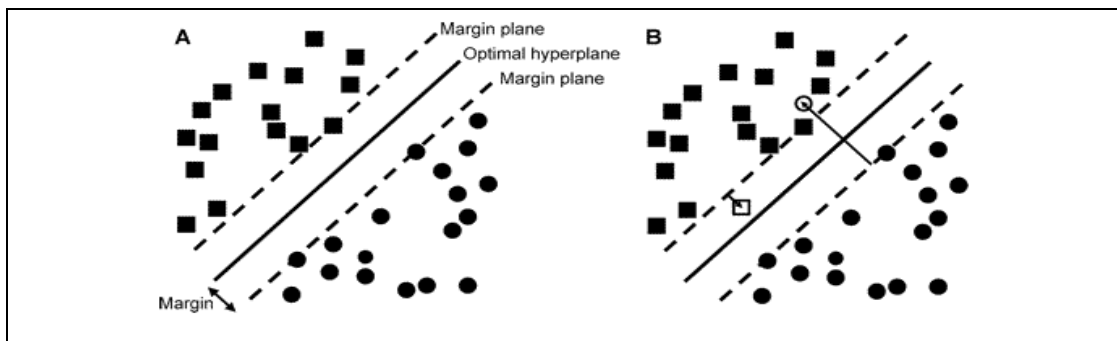


Figure 2.4: The separable hyper-plane of Vector Support Machine

This figure shows separating hyper-plane of the Support Vector Machine that maximises the hyper-plane between two sets of perfectly separable data instances, represented as circles and squares. (A) Optimal hyper-plane that perfectly separates the two classes of data instance. (B) Optimal soft margin hyper-plane, which tolerates some points (unfilled square and circle) on the “wrong” side of the appropriate margin plane (Jorissen and Gilson, 2005)

The Support Vector Machine (SVM) has successfully performed feature selection with wrapper methods in different fields of research such as bioinformatics, web mining and financial forecasting problems. In the bioinformatics field, Chiu et al. (2008) used SVM with wrapper to investigate the relevant genes that may help predict breast cancer and the speed with which the cancer will spread out around the human body. Their study revealed that breast cancer can be better predicted with a small number of the most significant genes (44 genes out of total of 403 genes in datasets) that are believed to be the most informative in the task of breast cancer classification. Their results show that the selected genes generated by SVM produced more accurate classification results than would have been achieved had all the genes in the database been used. In the web mining field, Abbasi et al. (2008) performed sentiment analysis of online content from webpage using wrapper and SVM algorithm. Their aim was to

Chapter Two: Literature Review

select the most relevant features that would help determine the type of information content (i.e. positive or negative information) from two web forums (Arabic n=13811 and English n= 12881). The wrapper method was able to select the most relevant information from the two forums (508 features from the English forum webpage and 338 features from the Arabic forum webpage). These selected features indicated higher performance accuracy than would have been achieved had all the features been taken. The relevant features were additionally found to be more useful in analysing more sophisticated document contents of the web forums.

SVM has proved successful in its ability to predict stock price directional movement in the financial market. Yang et al. (2002) used Support Vector Regression (SVR) to predict stock market volatility in the Hang Seng Index. Their findings revealed that using the standard deviation as a measure of stock volatility results in the best prediction in their model with minimum errors, which proves the ability of SVR to forecast stock market volatility. Lee (2009) developed a prediction model that combined a Support Vector Machine (SVM) with a hybrid feature selection method, namely F-score and Supported Sequential Forward Search (F_SSFS), with the aim of predicting stock market trend movements. Their proposed model was compared with a back-propagation neural network-based model (BPNN). Their experiment revealed that their model based on SVM and (F_SSFS) outperforms BPN in the problem of stock trend prediction. Kim (2003) used SVM to predict daily stock price movements in the Korean composite Stock Price Index (KOSPI). He used 12 technical indicators that served as the initial attributes for his experiment. His study aimed to investigate the predictive ability of SVM in financial forecasting by comparing it with back-propagation neural network (BPNN) and case-based reasoning (CBR). The results of his experiments showed that SVM outperformed BPN and CBR and proved itself to be a powerful technique in stock market prediction. Manish and Thenmozhi (2006) used the same technical indicators applied by Kim (2003), using SVM and random forest to predict daily movements of stock prices in the S&P CNX NIFTY Market Index of the National Stock Exchange. Their results were compared with those of the traditional discriminant and logit models and ANN. Experimental results proved that SVM outperforms Random Forest, neural networks and other traditional models. Huang et al. (2005) used a model that integrates SVM in combination with other classifiers that have proved successful in predicting the weekly movement direction of

Chapter Two: Literature Review

the NIKKEI 225 index. In order to evaluate the predictive ability of SVM, they evaluated its performance in comparison with those of linear discriminant analysis, quadratic discriminant analysis and Elman back-propagation neural networks. Their finding revealed that SVM outperforms the other classification methods. Hsu et al. (2009) developed two-stage architecture by integrating a self-organising map and support vector regression for stock price prediction. In their study they investigated the predictability of financial market movement in seven major stock market indices: the Nikkei 225 (NK), the All Ordinaries (AU), the Hang Seng (HS), the Straits Times (ST), the Taiwan Weighted (TW), the KOSPI (KO), and the Dow Jones (DJ). The results suggested that the two- stage architecture provides a promising alternative for stock price prediction. A recent study by Kara et al. (2011) compared the efficiency of two models, namely artificial neural networks (ANN) and support vector machines (SVM), in predicting stock price movements in the Istanbul Stock Exchange (ISE) National 100 index. They used ten technical indicator variables as inputs for their models. The experiments' results show that the ANN model performs better than the SVM model in stock prediction tasks.

2.6 The Role of Classifier in Feature Selection

Using different types of classifiers to perform feature selection results in different features being selected under each classifier. This is because each classifier has different biases and assumptions. Table 2.3 shows the three classifiers previously described with their biases and assumptions. As it can be seen from this table, each classifier has its own bias and assumptions, causing it to focus on different features when performing feature selection.

Table 2.3 shows that each classifier is based on different assumptions and possesses different biases, as well as the time required to perform feature selection will differ accordingly. For example, a complex classifier such as Support Vector Machine is likely to require more time to determine relevant features than a simple classifier, such as Naive Bayes Algorithm, would need. Moreover, classifiers with different biases and assumptions will select different relevant features, probably resulting in different levels of accuracy. In general, the amount of time and the level of accuracy are two major variables that are very important for determining the role of classifiers in feature selection.

Table 2.3: Biases and Assumptions of Different Classifiers

The Biases and Assumptions of Classifiers		
Classifier	Assumption	Biases
Naïve Bayes Classifier Algorithm (NB)	Assumes conditional independence in which each feature in the database holds a mutually exclusive state of being independent of other features.	Emphasises features that maximise conditional independence when building the graphical network.
Decision Tree Classifier (DT)	Assumes separation of instance variables based on predefined splitting criterion used by the classifier.	Emphasises features and attributes that are most relevant with respect to the target variables in the decision tree.
Support Vector Machine (SVM)	Assumes the data are linearly separable and follow the identical independent distribution (I.I.D)	Emphasises features that fall in a given class where maximum/optimal hyper-planes are found to widely separate the data instance of the two classes.

Source: Adopted with modification from Chrysostomou, (2008)

In summary, it is clear from the above discussion of related studies that a gap exists in the literature, because very few studies have been carried out to investigate the impact of stock micro-blogging forums on the prediction of stock markets, and very little is known about the predictive power of the collective sentiments of StockTwits regarding stock market performance. Moreover, most of the previous studies have focused on sentiment polarity (positive/negative polarity or emotional/mood states) for sentiment classification of online financial text. In addition, previous studies have focused on single classifiers to predict stock market movements. To the best of the researcher’s knowledge, no study has yet made a comparative analysis of performance accuracy of different algorithms designed to automatically detect sentiment from online financial text in stock forums. The gaps that will be addressed in this study are described in the next section.

Chapter Two: Literature Review

2.7 The Literature Gaps

This extensive review of relevant research on the predictive value of online investment forums in predicting financial market movements has identified several important research gaps.

The first gap in the literature concerns the lack of rigorous and in-depth analysis of the relationship between stock micro blogging features and financial market indicators. Although there has been a proliferation of research investigating the predictive power of online investment forums in predicting stock price movements in different contexts, there is a lack of research offering a precise analysis of financial market models to predict stock market behavioural movements. Most studies in the field have focused on simple lead-lag relationships (Antweiler and Frank, 2004b; Sprenger et al., 2014). There is a lack of research offering a precise analysis of the effect of sentiment in different market conditions (e.g. bull and bear markets). This research also takes a different approach to examining this predictive relationship by investigating the linear as well as the non-linear relationship between StockTwits and financial markets. Studying the non-linear model allows an investigation of the asymmetrical behaviour of investor sentiment by distinguishing between the bullish and bearish investors and how stock returns, volatility and trading volumes respond to the shift in investor sentiment (bullish and bearish shifts). Whilst the vast majority of empirical literature focuses on examining the relation between investor sentiment and stock market return in the form of linear regression frameworks, there is growing empirical evidence to suggest that stock returns may be better described by a model that allows for non-linear behaviour. For example, McMillan, (2005) suggests that stock returns might be better characterised by models that incorporate non-linear components of the explanatory variables. He argues that the interaction between arbitrageurs and noise traders makes it difficult to analyse stock returns using linear models. In addition to the non-linear model, a quantile regression model is also employed in this research to examine the relationship of the change in investor sentiment across different quantiles of returns distributions. Quantile regressions allow us to probe how the performances of stock markets affect the linkage between stock returns and investor sentiment.

Secondly, while the divergence of opinion among investors in the stock market and its prospective relationships with trading volume have been extensively

Chapter Two: Literature Review

addressed by scholarly research (Miller, 1977; Harrison and Kreps, 1978; Harris and Raviv, 1993; Diether and Scherbina, 2002; Basak, 2005), research on disagreement-volume relations has come to different conclusions on whether disagreement increases or decreases the trading volume. This research investigates this hypothesis but goes a step further to empirically investigate how divergence of opinion affects trading volumes considering different states of economies (the bull and bear markets).

Thirdly, apart from investigating the impact of disagreement on trading volumes, this thesis investigates the impact of investor sentiments on stock return and volatility. While a great deal of empirical literature in the field has investigated the role of investor sentiments and noise traders on stock return and volatility in the context of DSSW (1990) models, this study uses different data for investor sentiment measures. A number of sentiment measures have been used extensively in the literature, such as close-end fund discount (Lee et al., (LST) 1991; Neal and Wheatley, 1998; Brown and Cliff, 2004), household data (Kelly, 1997), country fund discount (Bodurtha et al., 1995), the index of investor sentiment change (Baker and Wurgler, 2006, 2007) and the Investor Intelligence survey (Kurov, 2010). This study, however, motivated by the DSSW (1990) model of noise traders, employs a relatively new form of data for investor sentiment extracted from stock micro blogging forums, the so-called "StockTwits". Although Sprenger et al. (2014) used the same measures of investor sentiment; their study did not investigate the impact of the noise traders as suggested by the DSSW (1990) model. DSSW (1990) argues that that the irrational noise traders with erroneous stochastic belief have the power to change prices and earn higher expected returns. The unpredicted beliefs of the noise trader can create risk in the price of assets, which prevents rational traders from aggressively betting against them. Despite the popularity of the DSSW models, which are becoming one of the significant theories in behavioural finance, relevant empirical studies are quite limited, especially with regard to the asymmetrical impact of investor sentiments on stock volatility. Lee et al. (2002) employed Investor Intelligence as a sentiment measure to test the four effects of noise trading on the price of risky assets. However, to the best of the researcher's knowledge, this is the first study in the field to explore the impact of noise traders, as suggested by the DSSW (1990), using relatively new data serving as a proxy for investor sentiments extracted from online stock forums, the so-called StockTwits.

Chapter Two: Literature Review

Fourthly, an extensive review of the literature indicates that although several approaches have been employed to perform sentiment classification for predicting stock markets, little emphasis has been placed on feature selection techniques. Previous research has focused on the sentiment polarity of online financial text in predicting stock market behaviour, such as positive/ negative polarity, emotional states and mood dimensions. To the best of the researcher's knowledge, no study has used feature selection in predicting investors' decisions in the stock market. There are many advantages associated with feature selection techniques. The first advantage is that it can potentially improve the classification accuracy by selecting the most relevant features, possibly resulting in a better understanding of the sentiment classification problem (Guyon and Elisseeff, 2003). The wrapper approach to feature selection can even provide a superior performance in terms of classification accuracy over the filter approach (Ruiz et al., 2006; Zheng and Zhang, 2008). With the wrapper approach, stock market prediction will be more accurate as it tends not only to investigate the single word effect but also takes into account the effect of pairs of words in predicting price movements and, hence, investor decisions. Another advantage is that, by focusing on highly relevant subset features, the class attributes will be narrowed down to key features (Gamon, 2004), thus reducing the risk of data over-fitting. Furthermore, with the removal of irrelevant features, feature selection techniques will have the ability to tackle the dimensionality reduction problem in the datasets. Given the high-dimensional market data used, stock market prediction usually involves a high computational cost and high risk of over-fitting (Lee, 2009). Thus, feature selection techniques have proved to be the most appropriate techniques in various stock prediction applications to overcome the high dimensionality and the risk of over-fitting.

Another important gap that has been identified is the fact that, despite the proliferation of research, which has used different classifier algorithms to investigate the predictive power of various online investing forums in predicting stock markets, most of the studies have focused merely on single classifiers to predict stock market movement (Schumaker and Chen, 2009; Mittermayer, 2004; Antweiler and Frank, 2004). There is a lack of research offering a comparative analysis of classification and performance accuracy in the use of multiple classifiers for sentiment detection in predicting stock markets from online text. The nature of each classifier along with its

Chapter Two: Literature Review

biases poses a problem in terms of affecting the classification accuracy. For example, a classifier with one bias may be more or less accurate than another classifier with a different type of bias. In the feature selection process, the use of different classifiers will result in different relevant features being selected under each classifier algorithm. This will lead to a different level of classification accuracy being reported for each classifier. This problem stimulates the need for the use of multiple classifiers for feature selection. However, little is known about the effect of using multiple classifiers for feature selection in predicting stock market indicators. The number of classifiers and the nature of these classifiers play a critical role in feature selection techniques that may affect the feature selection outcomes as well as the overall accuracy level of the features. Thus, it is of great importance to understand the effect of multiple classifiers for feature selection to improve the performance results of sentiment classification and improve market predictions as a result.

Thus, this study is a response to the above deficiencies. It therefore represents an early attempt to provide a more comprehensive examination and analysis of how sentiments extracted from stock micro blogging forums might successfully predict behavioural movements in financial markets while providing an in-depth analysis of these predictive relationships using rigorous econometrics modelling and machine learning techniques.

2.8 Chapter Summary

The aim of this chapter was to build a theoretical foundation for the empirical research through a critical review of the related literature. The chapter was divided into six main parts. The first part discussed popular theories that have been widely used in stock market prediction (EMH and RWT), and the theoretical and empirical success and challenges of these theories. Part two covered the behavioural finance theory and the importance of its complementary role in resolving the implications of traditional finance theory by highlighting the role of different traders in capital markets. The third part discussed the role played by noise traders in capital markets. Great emphasis was placed on two specific issues: how well investor sentiment can predict stock returns and whether disagreement is associated with more trades. The fourth part covered the issues related to social media, micro-blogging and virtual community forums while placing more emphasis on different forms of online stock

Chapter Two: Literature Review

forums (Internet Message Boards, Financial News Articles and Micro-blogging forums) and the underlying effects of each of these forums on financial markets. The fifth part offered definitions of text mining, explained various tasks of text mining, and elaborated in greater detail the feature selection task, which is the main interest of this research study. Different text mining techniques (Naïve Bias, Decision Tree and Support Vector Machine) and the role played by each of these classifiers in feature selection were covered in part six. In doing so, the research gaps were identified. It was revealed that there has been a lack of rigorous methodology in the analysis of financial indicators and a failure to implement multiple classifiers and feature selection tasks of text mining in predicting stock market behaviour on online stock forums.

Based on this critical review of the literature, the next chapter proposes a conceptual framework that investigates the predictive power of stock micro-blogging features in predicting different financial market indicators, employing different machine learning techniques for text classification.

CHAPTER THREE: CONCEPTUAL FRAMEWORK FOR

3.1 Introduction

In previous chapters, the research problem and aim of this thesis were defined. Then, relevant literature was critically reviewed and evaluated to identify important links between the problem and aim of this study. The basis for the theoretical development in this chapter is the delineation of the research problem and the review of literature in the previous chapters. Saunders et al.(2011) argue that a conceptual framework functions as a mechanism that enables researchers to connect the study with the existing body of knowledge on the research subject undertaken. It functions as a sensitising device helping the researcher “theorise or make logical sense of the research problem” (Sekaran, 2003, p. 87). Walsham (1995, p. 76) pointed out that “...An initial theoretical framework which takes account of previous knowledge... [would help in creating] a ...sensible theoretical basis to inform the topics and approach of the earlier empirical work”. A conceptual framework designs the key variables and constructs of the phenomenon being studied and the presumed correlations between them (Miles and Huberman, 1994). Moreover, Voss et al. (2002) argue that a graphical research framework design is considered an important starting point, as it provides a prior view of the general constructs and variables that a researcher intended to study alongside their expected relationships. Therefore, the framework in this chapter will serve as a guide for the investigation and presentation of a possible explanation for the phenomenon of the predictive ability of Stock Micro-blogging sentiments in forecasting stock price behavioural movements in financial markets.

This chapter consists of five main sections, including this introduction. Section 3.2 discusses the theoretical reasons for the assertion of this thesis on the existence of predictive power of Stock Micro-blogging sentiments in forecasting stock price movements in capital markets, which are utilised as a foundation for this research. Section 3.3 highlights the development of the proposed conceptual framework for this research and discusses the key sets of variables extracted from Stock Micro-blogging and the stock market accordingly, as previous relevant literature reveals that this is helpful for investigating the predictive power of Stock Micro-blogging sentiments in

Chapter Three: Conceptual Framework

forecasting stock price movements. Section 3.4 develops and addresses the research hypotheses that help to answer the research questions of this study. Section 3.5 provides a brief summary of this chapter.

3.2 Theoretical Foundation

This research study presents three theoretical discussions to support the assertion of this thesis on the existence of the predictive power of Stock Micro-blogging sentiments in forecasting stock price movements in the stock market.

While the research methods of this study are grounded in data mining, sentiment analysis and statistical computation, three reasons for our assertion stem from the finance literature. First, to address the topic of future stock price prediction, several theories are relevant. Several works have attempted to study stock market prediction while providing an answer to the common question - can stock prices really be predicted? The two theories presented in Chapter 2 (EMH and RWT) are the most relevant for answering such a question. Both theories present an interesting theoretical framework as the foundation for this study.

Second, the presence of the different types of investors in financial markets and the effect of their trading behaviour in influencing price changes are considered one of the reasons for our contention that stems from the behavioural finance literature. The two types of traders in financial markets (as already mentioned in chapter two) are: the “irrational noise trader” and “rational investors” or arbitrageur (Glosten and Milgrom, 1985; Kyle, 1985; Black, 1986; DeLong et al., 1990; Baberies and Thaler, 2003). Rational investors are those who make choices that are normatively acceptable and make sense (Baberies and Thaler, 2003). The presence of noise traders in financial markets can cause price levels and risk to deviate from expected levels even if all other traders are rational (De Long et al., 1990). Noise traders always participate in discussions and conversations related to financial information in capital markets. In the context of online investment forums, conversations among investors, including noise traders, involve making predictions, exchanging opinions, asking questions, sharing analyses and reporting financial information (Oh and Sheng, 2011). People tend to pay attention to ideas and facts that are reinforced by conversations (Hirshleifer, 2001). Friedman (1953) and Fama, (1965b) both argue that noise investors get together in the market with rational

Chapter Three: Conceptual Framework

arbitrageurs who trade against them and in the process drive prices close to fundamental values. This argument also supported by De Long et al. (1990) who first theorised the impact of noise traders by stating that noise traders with erroneous stochastic beliefs have the power to change prices and earn higher expected returns. The unpredicted beliefs of the noise trader can create risk in the price of the assets, thus preventing rational traders from aggressively betting against them (De Long et al., 1990). They argue that when sentiments of the noise traders are correlated with one another, they create risk.⁵ Furthermore, researchers such as Campbell and Kyle (1993) and Koski et al. (2004) assert that noise trading increases volatility and create risk termed “noise trader risk”. The role of noise traders has also been found in the context of financial community forums. Noise traders tend to engage in conversations regarding investment information by sharing investment opinion and analysis, asking frequent questions and making predictions. . Therefore, the ability of noise traders to cause price changes will also appear in online investment forums where the information and opinions are spread widely among investors through the investment communication platform channels (Zhang and Swanson, 2010). Stock micro-blogging is one of the platforms where the sentiment of noise traders plays an active role in information diffusion (Oh and Sheng, 2011).

The third reason for the above assertion is that the distinct features of stock micro-blogging provide great support for this study of the predictive ability of StockTwits sentiments in forecasting stock price movement in stock markets. The three distinct features are the high volume of message posts, the real-time message streams, and the succinctness that leads to the efficient diffusion of information in investment community forums (Java et al., 2007; Bollen et al., 2011). First, Micro-blogging services provide an easy way of sharing status messages either publicly or in a social network. The 140-character messages of Twitter posts can easily be read and followed by investors and other users, unlike long financial news articles whose length may cause investors to ignore parts of the articles. This results in posts that are to the point without much of the noise found in traditional blogs and articles. In addition, encouraging shorter posts saves the user’s time for more post contributions (Java et al., 2007). Second, the real-time conversations on stock micro-blogging forums produce up-to-date information on all stocks when compared to separate

⁵ Noise trader risk is a form of market risk associated with the trading decisions of noise traders, which cause price levels and volatility to diverge significantly from the fundamental or expected levels.

Chapter Three: Conceptual Framework

bulletin boards for each individual company, which results in outdated information if no new posts have recently been entered on that board.⁶ Claburn (2009) argues that as messages are generally being posted just before an event occurs, this implies that the forum contains real-time information that is important for making investment decisions. The real-time information transmitted in stock micro-blogging makes it a powerful decision support mechanism for investors. This contrasts with other forums where information is outdated as time passes with no new information being posted, thus resulting in deteriorating relevance and decreasing such forums' usefulness for planning and investment decision purposes (Ballou and Pazer, 1995). Third, StockTwits messages are highly suitable for this research study in terms of their relevance to the research topic. High-volume messages posted on Twitter's public timeline every day on a variety of topics make it difficult to extract relevant stock tweets of certain companies (e.g. Microsoft (\$MSFT) and Apple (\$AAPL)), as these companies are extensively discussed for purposes other than stock discussions (Ruiz et al., 2012). For example, Apple is a name that is frequently used for spamming purposes (e.g. "Win a free iPhone" scams). Since this study focuses on market conversations, StockTwits are highly relevant to this research topic.

3.3 Development of the Conceptual Framework

The conceptual framework focuses on the relationship between stock micro-blogging features and different financial indicator variables. The aim is to establish a framework for determining the predictive power of stock micro-blogging sentiments in predicting the behavioural movement of stock prices in capital markets. In order to develop an adequate framework, one needs to look at both StockTwits and financial market features. The StockTwits features consist of the following variables: message volume, bullishness (proxy for investor sentiment) and the level of agreement, while the market features consist of the following variables: return, trading volume and volatility. The conceptual framework encapsulated in Figure 3.1 is explicitly guided by highlighting the features extracted from both stock micro-blogs (StockTwits) and the stock market to study the correlation between those extracted features in order to investigate the predictive value of stock micro-blogs in predicting stock price movements in the stock market.

⁶Bulletin boards, sometimes referred to as message boards, are organised online forums enabling users to read and post information on specific firms and investment-related topics (Wysocki, 1998).

Chapter Three: Conceptual Framework

The following sections discuss the sets of key features, which have been revealed by previous literature as helpful for investigating the correlation between stock micro-blogging features and the related stock market features.

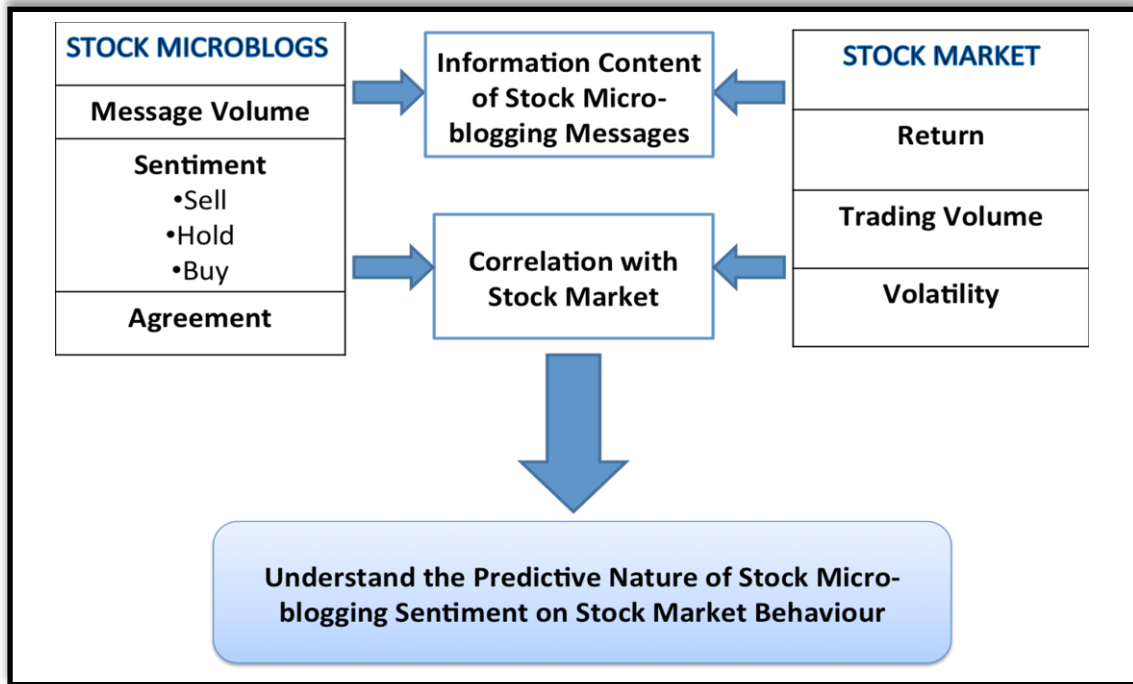


Figure 3.1: The conceptual framework

The following Section explains the conceptual framework and discusses the sets of variables used to investigate the relationship between the extracted features in order to address the effect of stock micro-blogging forums on the stock market.

3.4 Research Hypothesis

This section investigates the nature of the relationship and explores the linkage between the features extracted from Stock micro blogging (i.e. message volume, sentiment and level of agreement) with different stock market indicators (i.e. return, volatility and trading volumes) as indicated in Figure 3.1. The purpose of this section is to formulate the research hypotheses that need to be tested in order to investigate the predictive power of stock micro blogging forums in forecasting stock price behaviour in stock market while providing answers to the research questions. The following subsection addresses the development of the research hypotheses for this research thesis where the possible linkage of each of the stock micro blogging features are identified and described in relation to stock market variables.

Chapter Three: Conceptual Framework

3.4.1 Message Volume and Stock Market Features

Message volume is one of the main features of StockTwits, which indicates new information arriving in the market about particular discussed stocks. Therefore, in this research the volume of postings serves as a proxy for information arrival. In order to understand how message volume affects the stock market, it is important to begin by looking at the efficient market hypothesis (EMH), especially the semi-strong form. In an informationally perfect market, security prices reflect all publicly available information indicating investors' expectations of earning a return on their investments (Fama, 1965b; Reilly and Brown, 2009). However, information is seldom perfect in reality, and there is a requirement for economic agents (e.g. analysts' recommendations and financial professionals) to enhance information efficiency by incorporating their information into security prices (Grossman, 1976, 1995; Grossman and Stiglitz, 1980). Recent evidence suggests that qualitative information in particular is not instantly and fully reflected in stock prices. For example, Green (2006) examined 7,000 recommendation changes by 16 brokers and concluded that, after controlling for transaction costs, investors buying or selling following increases and decreases in prices could make abnormal returns of 1.02% and 1.5% for corresponding increases and decreases. In addition, market anomalies such as the price/earnings ratio suggest that investors can benefit by observing and analysing publicly available information (Tripathi, 2009). Deb (2012) provides evidence that value premiums in the Indian stock market are prevalent for both absolute returns (e.g., average returns and buy-and-hold returns) and risk-adjusted returns (e.g., Jensen's alpha, Treynor measure, Sharpe ratio). In addition, Tetlock et al. (2008) provide evidence that stock prices tend to under-react to the textual information contained in news articles.

The new branch of finance known as behavioural finance theory has also challenged the EMH. Behavioural finance theory suggests that proponents of the EMH base their arguments on the assumption that investors are rational. This indicates that investors always update their beliefs correctly at all times in response to the information provided to them. Baberies and Thaler (2003) contend that the behaviour of investors cannot be understood using the traditional framework presented by efficient market theorists. Behavioural theorists suggest a new framework based on the assumption that investors are not always rational. Despite the

Chapter Three: Conceptual Framework

arguments by behavioural theorists, Schwert (2003) argues that it is a mistake to attribute market anomalies to market inefficiency without considering the methodologies employed in the studies that document market anomalies. Even if markets are efficient, as suggested by the EMH, it is obvious that there might be temporal deviations from efficiency. This means that investors must correct for these deviations by attempting to take advantage of mispricing in the market. Message volume can play an important role in correcting for mispricing in the stock market. It can be argued that the message volume of stock micro-blogging can help predict the movement of stock prices.

StockTwits⁷ are characterised by three distinct features: succinctness, high volume and real time (Oh and Sheng, 2011). These features play a tremendous role in facilitating the diffusion of investment information among investors (Bollen et al., 2010; Bollen et al., 2011; Java et al., 2007). In StockTwits, for example, the content is considered brief and succinct in that users can only post short updates or postings that are less than or equal to 140 characters in length (Oh and Sheng, 2011). The brief nature of the messages helps minimise noise and improves the relevance of the information contained in the messages. In addition, the time taken to transmit the messages is low, which results in a high volume of postings. Furthermore, messages are generally being posted just as an event occurs, which means that the forum contains real-time information that is important for making investment decisions (Claburn, 2009). Stock micro-blogging is considered a means of gaining access to real-time information because, as time passes, the relevance deteriorates and it becomes less useful for planning and decision-making purposes (Ballou and Pazer, 1995).

Relevant publicly available information, such as company press releases, price/earnings ratios, analyst recommendations and earnings announcements, tends to be sporadic and less frequent. This means that the continuous streaming of micro-blogs serves as a new frequent and constant source of information to investors that would otherwise be unavailable. Based on the foregoing discussion, one can argue that message volume, real time and succinctness have an impact on micro-blogging sentiments and, thus, on the movement of stock prices. The following subsection

⁷StockTwits and stock micro blogging will be used interchangeably throughout the text in this thesis.

Chapter Three: Conceptual Framework

formulates the hypothesis testing of the relationship between message volume and the three financial indicator variables as shown in the following figure.

As it can be seen from the Figure 3.2, while message volume tends to directly affect the trading volume, return and volatility seem to be affected indirectly by message volume through its effect on trading volume (as indicated by the dotted line from trading volume to return and volatility). Hence, trading volume is operating as an intermediary variable to reflect the effect of message volume on both return and volatility. However, message volume also seems to have a direct impact on volatility, as shown by the solid line directed from message volume to volatility (Figure 3.2.)

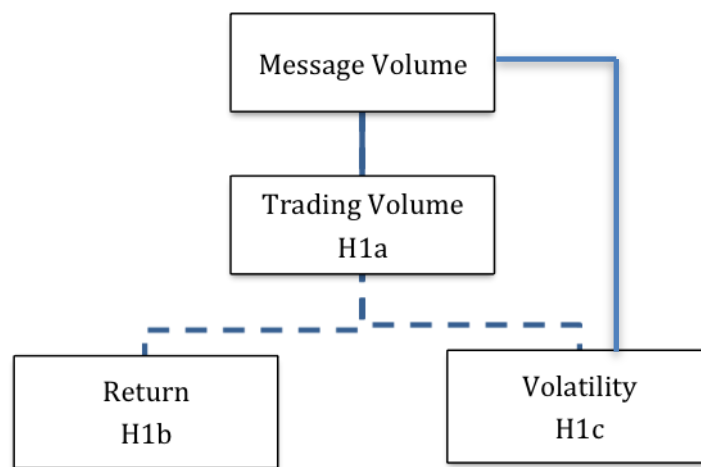


Figure 3.2: the relationship between message volume and stock market variables (trading volume, return and volatility).

The following subsections provide detailed explanations of the prospective relationships between message volumes and financial market indicators such as trading volume, returns and volatility.

▪ Message Volume and Trading Volume

Generally, message volume can affect trading volume, volatility and stock market return. With regard to trading volume, as more messages are posted on the forum, more and more investors start developing sentiments about the prices of that particular stock. As a result, the likelihood of trading increases. Therefore, one would expect to observe a positive relationship between message volumes and trading volumes. It is obvious that people are motivated to post information about the stocks in which they are trading (Van, 2003). This is consistent with empirical evidence such

Chapter Three: Conceptual Framework

as that obtained by Antweiler and Frank (2004b) who provide evidence that trading volume has a positive link with message volume. However, Das et al. (2005) argue that postings in online forums reflect the activity of retail investors who trade in small volumes on a daily basis rather than large institutional investors who trade in bulk. This indicates that an increase in message volume should not significantly affect trading volume. However, if one goes beyond the direct relationship between trading volume and message volume, higher message volume will motivate even "lurkers" to trade (Sprenger et al., 2014). A lurker is a member of an online forum who observes but does not actively participate or post. Therefore, this research study suggests that communication among participants in the market motivates trading by the kind of investors who may decide to trade when they become aware that other traders share a similar view. Consequently, message volume in stock micro-blogs should reflect the communication of investors. Therefore, this study will consider the following hypothesis:

H1a: Message volume in stock micro-blogging forums has a positive impact on trading volume.

▪ Message Volume and Stock Return

Message-posting volume is expected to predict stock returns implicitly through its effect on volume of trade in the security market. Trading volume plays a tremendous role in determining assets return. The relationship between assets return and trading volume is essential for demonstrating and understanding operational efficiency and information dynamics in the stock market. Some studies have focused exclusively on the impact of trading volume on stock returns. Chen (2012) empirically tests the long-run equilibrium relationship between stock returns and trading volume by focusing on the impact of stock market dynamics (unobservable or latent variables) on stock returns and trading volume. Using monthly stock prices and volume data for the S&P 500 Index in bear and bull markets, he provides evidence of a significant and positive link between trading volume and stock return across bull and bear markets. Trading volume is also found to significantly affect the cross-section of stock returns across bull and bear markets. Using two sub-periods, the study observed that the evidence is consistent across both sub-periods, which indicates that the results cannot be attributed to chance. Lamoureux and Lastrapes

Chapter Three: Conceptual Framework

(1990) use daily trading volume data (collected from S&P's daily stock price record) as a proxy for information arrivals time and have found that trading volume has explanatory power in explaining daily variations in returns on actively traded stocks with listed options traded in the Chicago Board Options Exchange (CBOE). Hiemstra and Jones (1994) studied the linear and non-linear Granger causality tests to examine the dynamic relations between daily returns on the Dow Jones and percentage changes of trading volume of stocks traded on the New York Stock Exchange. Their findings reveal significant bidirectional non-linear causality between returns and trading volume. Despite the evidence in support of a potential impact of trading volume on stock returns, some studies have observed that there is no significant impact of trading volume on stock price changes. Lee and Rui (2002), for example, observe that trading volume has no significant effect on stock returns based on data from the New York Stock Exchange (NYSE), the London Stock Exchange (LSE) and the Tokyo Stock Exchange (TSE). The evidence on the long-run equilibrium relationship between trading volume and stock returns becomes even more confusing given that some studies have observed that it is returns that affect trading volume and not vice versa.

While the relationship between trading volume and stock returns has been explicitly demonstrated, and since the trading volume has been found to be directly affected by message volume, one would expect message volume to have an impact on stock returns. An increase in message volume is an indication that new information is arriving in the market. An increase (decrease) in message volume would be interpreted as a bullish (bearish) attitude to a particular discussed stock. According to Dewally (2003), most of the messages in stock forums often represent buy signals, and he shows that an increase in message volume can be interpreted as bullishness. As investors (noise traders) became more bullish (bearish) about a particular asset, their demand to purchase (sell) that asset increased (DSSW, 1990). Therefore, when the demand of noise traders increased (decreased) relative to their average change in sentiment of being more bullish (bearish), they would expect a higher (lower) return relative to the market risk bearing. Hence, the changes in the volume of messages as a result of sentiment change would have an impact on stock returns. While the study by Antweiler and Frank (2004) provided evidence of a negative relationship between message volume and stock returns, Sabherwal et al. (2008) obtained contradictory

Chapter Three: Conceptual Framework

results and concluded that the most actively discussed stocks tend to exhibit significantly positive abnormal returns on the next trading day. Wysocki (1998) finds weak evidence of the explanatory power of an increase in message volume for positive next-day abnormal returns.

Based on the foregoing, this study will consider the following hypothesis:

H1b: Increases in message volume in stock micro-blogging forums are associated with higher stock returns.

▪ **Message Volume and Stock Return Volatility**

Message volume can also affect stock return volatility. As earlier noted, message volume can affect trading volume. Empirical evidence suggests that trading volume can influence stock return volatility. This means that message volume can affect the volatility of stocks through the indirect impact on trading volume. An earlier study by French and Roll (1986) argued that volatility is likely to be higher during the trading day than it would have been in non-trading hours due to differences in the flow of information. This is also true since messages are associated with the flow of information; therefore, the increase in message volume may affect stock return volatility.

Previous studies have tried to understand the link between trading volume, volatility and trading volume, and stock price performance (Karpoff, 1986; Pyun et al., 2001; Huang and Yang, 2001; Bohl and Henke, 2003). These studies have suggested two main hypotheses: the Mixture of Distribution Hypothesis (MDH) and the Sequential Arrival Information Hypothesis. The MDH, for example, suggests that stock return volatility and trading volume are determined by the same rate of information arrival or news process. This means that trading volume and volatility are likely to have a long-run equilibrium relationship (Clark, 1973). A number of studies have investigated the MDH and arrived at different findings. Pyun et al. (2000), for example, provide evidence in support of the MDH based on data from the Korean stock market. Similarly, Bohl and Henke (2003) findings agree with the MDH using data from the Polish stock market (Bohl and Henke, 2003). Lucey (2005), however, observes contrary findings based on an analysis of the Irish stock market. Raganathan and Pecker (1997) observe a positive relationship between trading volume and volatility based on an analysis of the Australian stock market. This evidence is

Chapter Three: Conceptual Framework

consistent with the MDH in that trading volume is likely to be determined by the same latent or unobservable variable. Most of the studies on the relationship between trading volume and stock return volatility and, thus, the MDH are based on a GARCH (1,1) specification. The GARCH (1,1) specification suggests that volatility tends to persist over time (Andersen et al., 2001, 2003; Baillie, 1996). Lamoureux and Lastrapes (1990) investigate the residual effects of the extended GARCH (1,1) model to account for the effects of trading volume. Chen et al. (2001) observe that extending the conventional GARCH (1,1) specification does not help in accounting for the persistence in volatility. Huang and Yang (2001) examined the MDH using data from the Taiwanese stock market. Their study is different from other studies in that it employs high-frequency data (5-minute stock returns) from the Taiwan Stock Index (TSI). The evidence suggests that the persistence in volatility does not disappear when trading volume is included as an additional factor in the GARCH (1,1) model. Most studies suggest the presence of a linear relationship between volatility and trading volume and fail to find a significant link between the two variables. Huang and Yang (2001), however, suggest the presence of a non-linear relationship between trading volume and volatility by observing a distinctive U-shaped pattern.

The foregoing evidence suggests stock return volatility is determined by trading volume. Given that message volume can impact trading volume, where an increase in message volume results in an increase in trading volume, this in effect suggests that message volume has an indirect impact on volatility through its direct impact on trading volume. The DSSW (1990) model suggested that the trading activities of noise traders may be correlated, thus causing biased noise traders to follow each other in selling or buying assets just as others are selling or buying. This leads to an increase in volatility as a result of the unpredictability of noise traders' beliefs, in turn creating a risk that deters arbitrageurs from dealing against them (e.g., Black, 1986; De Long et al., 1990). Message volume (proxied for information arrival) has been found to have a direct effect on stock return volatility. For example, Danthine and Moresi (1993) argue that the arrival of more information on the market decreases market volatility as this information puts rational investors into a better position to counteract the actions of noise traders. Antweiler and Frank's (2004) study of internet message boards provides empirical evidence of the predictive power of message volume on stock return volatility. Brown (1999) provides evidence that the

Chapter Three: Conceptual Framework

action of noise traders induces volatility. Koski et al. (2004) argue that noise trading proxied by message volume results in increased volatility in returns. Their study suggests that the vast majority of message board participants are considered day traders (noise traders) who trade on noise. Hence, given that a large proportion of stock micro-blogging forum participants are day traders, whose trading activities are expressed through message volume, the following hypothesis is derived:

H1c: Message volume in stock micro-blogging forums has a positive impact on stock return volatility.

3.4.2 Investor Sentiment and Stock Market Features

Investor sentiment, broadly speaking, refers to excessive optimism or pessimism about specific security prices (Lee et al., 1991; Antoniou et al., 2013). Another possible definition of investor sentiment is stated by Baker and Wurgler (2006) as “the propensity to speculate”. According to this definition, the relative demand for speculative investments is driven by sentiment. The propensity to speculate therefore has an effect on the performance of stock prices as well as on the volatility of stock returns. One factor that determines investor sentiment is subjectivity in the valuation of securities. For example, a less profitable potential growth company with limited earnings history will force less-informed investors to choose with an equal amount of probability a wide range of valuations depending on their sentiments (Baker and Wurgler, 2006). During bubble periods, when sentiment is high, investment bankers tend to argue in favour of high valuations of stocks. On the other hand, changes in investor sentiment tend not to significantly affect the valuation of firms with long earnings histories, tangible assets and stable dividends.

A large body of literature demonstrates that investor sentiment influences stock price behaviour, implying that financial economists should not only be aware of whether sentiments affect prices but should also shed light on and reinforce the extent to which investor sentiment impacts the stock market. After the theoretical work of De Long et al. (1990, 1991), many empirical research studies have continued to investigate evidence that sentiment, which reflects differences in the opinions of investors, has an influence on the future prospects of the stock market. For example, Shiller et al. (1996) adopted a direct approach to collect market sentiments and opinions from retail investors in both Japan and the United States by sending a

Chapter Three: Conceptual Framework

number of mail surveys to elicit their direct opinions and sentiments. A study conducted recently by Yu and Yuan (2011) examines the effect of investor sentiment on the mean-variance relation. They documented that, at the aggregate market level, sentiments significantly influence the mean variance trade-off and therefore proposed a model of stock price and risk-return relations that incorporated investors' collective sentiment. The findings of the above-mentioned studies question the underlying assumptions of the Efficient Market Hypothesis (EMH), which will be intensively discussed in the following section.

The findings of the above-mentioned studies have demonstrated the importance of investor sentiment in affecting stock price behaviour in capital markets. The following subsections examine the relationship between the investor sentiment, stock return, volatility and trading volume. Figure 3.3 below shows the linkage between investor sentiments with three related stock market variables (trading volume, return and volatility).

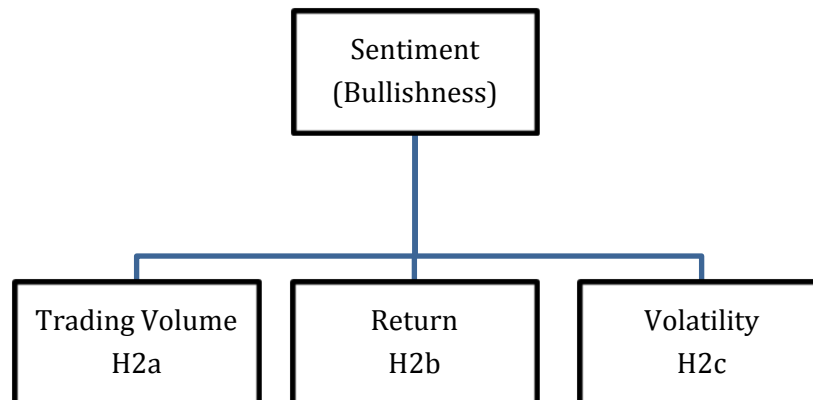


Figure 3.3: The relationship between investor sentiment and stock market variables (trading volume, return and volatility).

▪ Investors' Sentiment and Trading Volume

Economists of behavioural finance argue that behavioural changes in investor sentiment (optimism/pessimism) in assessing price valuation of assets (DSSW, 1990) have an impact on trading volume. Baker and Wurgler (2006) suggest that investors exhibiting excessive optimism or pessimism have the propensity to speculate through their excessive selling and buying activities in the financial market. Individual

Chapter Three: Conceptual Framework

investors who trade on noise with no fundamental information (Black, 1986) are more likely to buy and sell stocks in concert (Kumar and Lee, 2006). Therefore, the change in investor sentiment will have a great impact on trading volume via the changes in the trading activities. This prospective relationship is better addressed by highlighting the underlying assumptions of the Efficient Market Hypothesis (EMH). The first underlying assumption of the EMH is that all investors are rational. Behavioural finance theory, however, criticises this assumption and suggests that the existence of two types of investors (rational and irrational) along with their behavioural differences (sentiments) will have a direct impact on the trading volume of the traded securities in stock markets. Rational investors (arbitrageurs) always act rationally by correcting their beliefs directly when information arrives on the market and they are said to follow Bayesian beliefs. Noise traders, in contrast, are those who hold random beliefs and trade on noise as if new information has appeared in the market (Black, 1986; De Long et al., 1990). The trading activities of both noise traders and arbitrageurs greatly affects prices, which in turn affects investors' trading decisions in selling and buying assets in the financial market.

DSSW (1990) highlighted several characteristics of noise traders in their model. First, noise traders value financial assets based on noise and they represent information that has no fundamental components. Second, they depend on their psychological biases (overreacting or under-reacting) in processing information and forecasting stock returns. Third, they recognise risk incorrectly in the market. Therefore, noise traders will cause prices to fluctuate and depart from their fundamental values. On the other hand, rational arbitrageurs, who have Bayesian beliefs, will trade in a way that brings prices back to their fundamentals by buying securities when noise traders bring the prices down and selling them when they raise the price. These differences in behaviours of the two types of traders in the market demonstrate the effects of investors' sentiments in influencing stock price behaviours and consequently influencing the trading volume of the traded securities in the financial market. Since trading volume can be explained in terms of market liquidity, many researchers have studied the relationship between investor sentiments and stock market liquidity. For example, Baker and Stein (2004) and Liu (2015) show that investor sentiments are positively correlated with stock market liquidity. In addition,

Chapter Three: Conceptual Framework

Hong and Stein (2007) demonstrate that trading volume would be a good indicator of investment sentiments in the market.

The second assumption underlying EMH that is called into question is that investors' errors are not correlated since the investors are trading randomly and cancelling one another out. As a result, the advocates of EMH argue that the effect of noise traders is not significant and they do not change the fundamental value of the security; the market will always be efficient. However, researchers argue that noise trader sentiments are correlated with one another and will result in an increase in volatility in asset returns (Koski et al., 2004), which in turn affects the volume of trade in those assets. Other researchers argue that investments could be a social activity in a way that all investors encounter the same degree of risk exposure (Shiller et al., 1984). In fact, the effect of social influence on investing cannot be ignored. Investors are normally subject to information exposure, most often to rumours or noise, made available by their peers, financial professionals, family, friends and neighbours in daily conversations or in formal talks. According to the theory of DeMarzo (2003), people communicate in order to influence one another. Hong et al. (2005) and Duflo and Saez (2002) initially explored this idea when they argued that talk influences the actions of others.

Shiller et al. (1984) demonstrate the importance of social influences in affecting the behaviour of investors and thus affecting stock price behaviour as a result. They argue: "Investing in speculative assets is a social activity. Investors spend a substantial part of their leisure time discussing investments ideas, reading about investments, or gossiping about others' successes or failures in investing. It is thus plausible that investors' behaviour (and hence prices of speculative assets) would be influenced by social movements...Most of those who buy and sell in speculative markets seem to take it for granted that social movements significantly influence the behaviour of prices..." (1984, p. 457). Such social influence alters investors' behaviour in two aspects. First, the psychological change in investors' attitude in the market represented by their sentiments will have a great effect in altering the investors' decision-making and driving asset prices. Baker and Wurgler (2006) argue that, in practice, irrational investors trade more frequently and therefore add more liquidity when they are optimistic and betting on rising stocks rather than pessimistic and betting on falling stocks. This implies that if investors are optimistic about a particular

Chapter Three: Conceptual Framework

stock, they are more likely to hold the stock for a long time, which is a good signal to other investors to demand more of that particular stock and therefore results in upward trends of stock prices. In contrast, if investors are bearish (pessimistic) they are likely to stay short and will tend to sell that stock, thus giving a bad signal to other investors not to buy that particular stock and driving asset prices down as a result. Second, the trading transactions of individual investors in the stock market may be correlated, suggesting that this systematic correlation can cause the stock return to move in a lock-step. When investors follow each other as closely as possible by trading (selling and buying) in a concerted manner, they are said to be walking in a lock-step, which will result in a lock-step movement of stock returns of traded stocks. Kumar and Lee (2006) find that the trades of individual investors are systematically correlated when they buy or sell stocks in concert.

Retail investors are considered the least informed participants in the market (Hirschleifer and Teoh, 2003). Previous studies propose that retail investors are uninformed and undergo various behavioural biases (e.g., Odean, 1998a and b; Barber and Odean 2000; Benartzi and Thaler, 2001). Evidence shows that retail investors pay a significant price for trading actively (Sprenger et al., 2014). Despite this evidence, Mizrach and Weerts (2009) provide evidence that 55 per cent of retail investors in an Internet chat room actually made a positive return after adjusting for transaction costs. From a theoretical point of view, sophisticated investors with less trading capacity often find it difficult to fully exploit their trading capacity. This group of investors often have residual private information. These sophisticated investors are often motivated to spread the private information, thereby providing followers with reliable information. Followers trade on the advice of sophisticated investors. This therefore enables both sophisticated and retail investors to fully capture the value of private information (Bommel, 2003). Despite this evidence, there is an opportunity to spread false rumours in micro-blogging forums, which results in moral hazard (Bommel, 2003). Moral hazard can cause followers to ignore rumours altogether, thus resulting in a decline in trading volume. This is consistent with the strong-form of EMH which states that stock prices reflect all information, both public and private (Reilly and Brown, 2009).

In addition, Dewally (2003) observes that stocks that were recommended on message boards were stocks that had previously performed well, suggesting that

Chapter Three: Conceptual Framework

market participants follow a naive momentum strategy. Das and Chen (2007) argue that message bullishness and market returns exhibit only a contemporaneous relationship. Despite the above evidence, Hirshleifer and Teoh (2003) provide evidence that informationally equivalent disclosures can affect the perceptions of investors in different ways owing to limited attention and processing ability. For example, Barber and Odean (2008) suggest that retail investors prefer investing in stocks only if they have been made aware of these stocks through the news media. In addition, Ng and Wu (2006) suggest that investors tend to be motivated by word of mouth. In a micro-blogging forum, for example, Mizrack and Weerts (2009) observed that investors were more likely to follow the direction of their peers following a recent post on the same stock.

The foregoing evidence suggests investor sentiments and the behavioural/psychological differences of investors and their trading activities have an impact on trading volume of the traded assets in the stock market that can be explained in terms of the stock market liquidity through the changes in investors' selling and buying decisions in the stock market. This leads to the following hypothesis:

H2a: Investor sentiment derived from stock micro-blogs results in an increase intrading volume.

▪ Investor Sentiments and Stock Market Return

Having found that investors' sentiments have an impact on trading decisions, it becomes important to ask whether this subsequently translates into an effect on stock market returns. A well-known set of studies on sentiment and stock market returns emerged in the 1980s. Over the last few decades, a growing body of literature has empirically investigated the role of investor sentiment in stock markets and has provided significant evidence that investor sentiments are closely correlated to stock prices in financial markets. More specifically, evidence from the behavioural finance literature has proved the predictive ability of investor sentiments and the trading activity of noise traders in forecasting stock market returns (Shleifer and Summers, 1990; DSSW, 1990; Campbell and Kyle, 1993; Kelly, 1997). Previous studies focused on the time-series relationship between investor sentiment and stock price and

Chapter Three: Conceptual Framework

suggested that stock return is one of the most important factors affected by sentiment (Fisher and Statman, 2000; Brown and Cliff, 2004; Baker and Wurgler, 2006, 2007). For example, Fisher and Statman (2000) argue that sentiments of both small and large investors are reliable contrary indicators for predicting future S&P 500 index returns. In spite of the statistical significance of the relationship between individual investor sentiment and future S&P 500 returns, this relationship is found to be negative, implying that a higher individual investor sentiments index is followed by subsequent low future return of the S&P 500. An extended study by Fisher and Statman (2003) has examined whether the consumer confidence index might be used as a proxy for investor sentiment and predicted stock market return. Their result suggests that a high consumer index is associated with statistically significant increases in the bullish sentiments of individual investors about the stock market. Their findings are consistent with their earlier work, which found that higher consumer confidence is associated with a low S&P 500 index return.

A similar approach was adopted by Charoenrook (2005) who found that changes in consumer sentiment are economic and statistical predictors of stock returns in the market. Findings show a positive correlation between changes in consumer confidence and contemporaneous excess market returns, but they are negatively related to future excess market returns. Consistently, Brown and Cliff (2004) show a strong positive relation between changes in consumer confidence and contemporaneous stock return while showing that sentiment has little predictive power in predicting near term-future returns. Baker and Wurgler (2006, 2007) observe that poorly capitalised, young, smaller, unprofitable, highly volatile, non-dividend-paying, speculative companies with great potential for growth tend to have much higher returns during low investor sentiment. However, when sentiment is high, this category of stock tends to earn relatively low returns. Safer investments such as bonds and safer stocks are less driven by sentiment and, as such, their returns are less susceptible to investor sentiment. Baker et al. (2012) observe that, during the recent global financial crisis, the Morgan Stanley Capital International (MSCI) World Index of industrialised economies, emerging markets and the Chinese local market index dropped by 50%, 66% and 71%, respectively. The large declines in stock market indices indicate that investor sentiment contributed significantly to the movement of stock returns during the global financial crisis. They also show that financial crises

Chapter Three: Conceptual Framework

sway differently across different countries and regions. The foregoing evidence also shows that the behaviour of stock returns is determined by both local and global sentiments. Sentiment therefore contains distinct and strong explanatory power in determining assets returns.

Some studies have examined the impact of investor sentiment on the stock market using panel data methods (Schmeling, 2009). The advantage of using a panel data model is that the number of available observations increases tremendously, thus allowing the use of more informative data in the analysis. Panel data methods can therefore enhance the statistical power. For example, Baker and Wurgler (2006) suggest that the impact of sentiment is high when panel regressions are used. The literature on investor sentiment focuses mainly on the impact of sentiment on stock valuation using firm-level or aggregate data. Few studies have analysed how sentiment affects firm value. Assuming that investors buy more stocks in bull markets than in bear markets, this study suggests that different industry/country valuations are affected by different degrees of investor sentiment. For example, Kaplanski and Levy (2010) use panel data to determine the industries that are most affected by sentiment while Schmeling (2009) makes use of panel data to illustrate the types of countries that are most affected by sentiment.

Most studies investigate sentiment using linear models. However, McMillan (2005) suggests that stock returns can be better characterised by models that incorporate non-linear components of the explanatory variables. The study suggests that the interaction of arbitrageurs and noise traders makes it difficult to analyse stock returns using linear models. This indicates that their investor sentiment has asymmetric effects on the behaviour of stock returns and these effects can only be adequately accounted for by making use of non-linear models. Chung et al. (2012), for example, provide evidence that investors do not constantly deal with uncertainty as suggested by earlier evidence. The assumption of constant uncertainty is a result of the use of linear models. This assumption is invalidated when non-linear models are used. Several empirical studies provided evidence that the impact of investor sentiment on stock prices is asymmetric (e.g., Brown and Cliff, 2005; Gervais and Odean, 2001; Wang, 2001; Hong et al., 2000). That is, markets react differently to the various levels of investor sentiment; in particular, market performance during periods of growth (or recessions) induces an optimistic (or pessimistic) attitude in investors.

Chapter Three: Conceptual Framework

Indeed, DeBondt (1993) revealed that increased bullishness can be expected after a market rise compared with increased bearishness after a market fall, confirming the 'positive feedback traders' hypothesis. Verma and Verma (2007) have provided further evidence of the existence of asymmetric effects of the stock market on investor sentiment by emphasising the magnitude of effects of bullish (or bearish) sentiment in different states of the market: growth and decline. Their findings revealed that variations in the stock market may have stronger effects on bullish sentiment in a period of growth compared to market variation effects on bearish sentiments in a period of decline.

From the above discussion, evidence shows that investor sentiment has a predictive power in anticipating stock returns in the capital market. While sentiment-returns relations have long been addressed using linear models and have been proved significant in most studies, the incorporation of non-linear components has explored unviable aspects embodied in such relations. Non-linear models allow us to investigate the asymmetrical behaviour of investor sentiment and stock returns by differentiating between bullish and bearish sentiments. Moreover, investigating the impact of sentiment-returns relations in different states of economies, such as bull and bear markets, may also provide greater insights into how such relations might be affected in different regimes of the market. Therefore, the following hypothesis is proposed:

H2b: Investor sentiment derived from stock micro-blogs results in an increase in stock market return

▪ Investor Sentiments and Stock Return Volatility

The behavioural finance approach suggests that the behaviour of noise traders plays an important role in influencing stock price behaviour. In contrast to EMH, the biased beliefs of noise traders create risk that prevents arbitrageurs from correcting stock prices back to their fundamentals (Black, 1986; De Long et al., 1990), thus leading to an increase in stock price volatility. The unpredictability of noise trader beliefs (or so-called sentiment) results in biased encouragement to either sell or buy a particular stock traded on the stock market.

Chapter Three: Conceptual Framework

Apart from investigating the effect of investors' sentiments on stock market return, this thesis also examines whether investor sentiments have an impact on the volatility of stock returns. The price risk increases as noise trader beliefs become more variable (De Long et al., 1990). The DSSW model suggests that when the noise trader beliefs are correlated with one another, the presence of these noise traders in the market will create risk which was previously unseen, resulting in increased market volatility of security prices (De Long et al., 1990). The model further shows that the amount of market volatility depends heavily on the proportion of noise traders' presence in the market; i.e. the more noise traders present in the market, the greater the risk they pose to a particular security.

Despite the DSSW theories that investor sentiment and noise trading can influence stock return volatility, very little empirical evidence has been obtained to address this kind of relationship, in contrast to the well recognised evidence addressing the relationship between investor sentiment and stock market return. In spite of other supportive evidence in this line of research, it mainly focuses on the US market. A study conducted by Brown (1999) provides evidence that individual investor sentiment is associated with high volatility. He argues for the underlying assumption of EMH that the action of irrational investors not only influences asset prices (moving prices away from their fundamental values) but also incurs additional volatility. A number of research papers such as Black (1986), DeLong et al. (1990), and Campbell and Kyle (1993) have been testing the hypothesis that noise trading increases volatility. Koski et al. (2004) found supportive evidence to confirm the hypothesis that noise trading results in increased volatility and also confirmed the strong reverse causality test on that hypothesis. The Investors Intelligence sentiment index has been used by Lee et al. (2002) to study the relationship between this sentiment index and stock market returns and volatility. Their findings reveal that the bullish and bearish sentiments of investors have a notable effect on both returns and volatility as the changes in sentiment result in downward (or upward) market volatility and high (or low) excess return.

The foregoing suggests that sentiment can significantly affect the behaviour of stock returns. The noise trader literature contends that sentiments of noise traders are correlated with one another and, as such, result in an increase in stock return volatility (Koski et al., 2004). Noise traders can influence the behaviour of stock prices when

Chapter Three: Conceptual Framework

information is shared among investors through social media such as Yahoo Finance (Zhang and Swanson, 2010). Stock micro-blogging plays a tremendous role in providing a mechanism through which investors can share and spread information quickly through the continuous streaming of information using online media. In particular, sentiment, proxied by opinions, has a significant effect on the diffusion of information through online media. Developing more sentiment in such forums will cause traders to trade more actively and follow the trade direction (i.e., buy vs. sell) of their relative peers in the market, which in turn causes the price to fluctuate. This indicates that stock micro-blogging can have a significant impact on diffusing behavioural investor sentiments, which affect the behavioural movement of stock prices in the capital market. This study therefore argues that sentiment derived from stock micro-blogging can predict stock prices' volatility movement. This leads to the following hypothesis:

H2c: Investor sentiment derived from stock micro-blogs results in an increase in stock return volatility

3.4.3 Agreement and Stock Market Features

The level of agreement among investors in the StockTwits forum can have an impact on the stock market. When there is agreement on a micro-blogging forum, there will be a tendency for the stock prices to maintain a constant trend in a particular direction. The role of disagreement in affecting stock price behaviour in the capital market is one of the harder relations to predict in empirical finance. The price optimism model suggested by Miller (1977) implies that a high level of disagreement among traders would cause the market price of the stock to diverge from the fundamentals and to be relatively higher than its intrinsic value; this would cause investors with optimistic beliefs to trade even more, thus lowering expected returns. However, this upward bias in price values of securities does not exist in some models, such as those accounting for the rational component behaviour (Diamond and Verrecchia, 1987; Hong and Stein, 2003). Diamond and Verrecchia (1987) developed a model that incorporated perfectly rational investors with unlimited computational ability who immediately updated their beliefs to adjust the security price, reflecting all publicly available information. Hong and Stein's (2003) findings reveal that the presence of rational arbitrageurs can eliminate the risk of misperceptions caused by

Chapter Three: Conceptual Framework

noise traders in the capital market. However, their findings are called into question when the concept of limits on arbitrageurs was first theorised by DSSW (1990), who argued that the unpredictability of noise traders' beliefs creates a risk that prevent rational arbitrageurs from betting against them. Several later studies (Shleifer and Vishny, 1997; Gromb and Vayanos, 2002; Chen et al., 2002) support the DSSW model and provide compelling theoretical explanations of why arbitrageurs may fail to close the arbitrage opportunity.

The above studies have provided evidence that divergence of traders' opinions contains value-relevant information that is not yet reflected in security prices in the capital market. The effect of disagreement on trading volume, return and volatility and the linkages between them are shown in Figure 3.4 while the detailed explanation and the hypothesis formulation of each prospective relation is also provided in the forthcoming subsections.

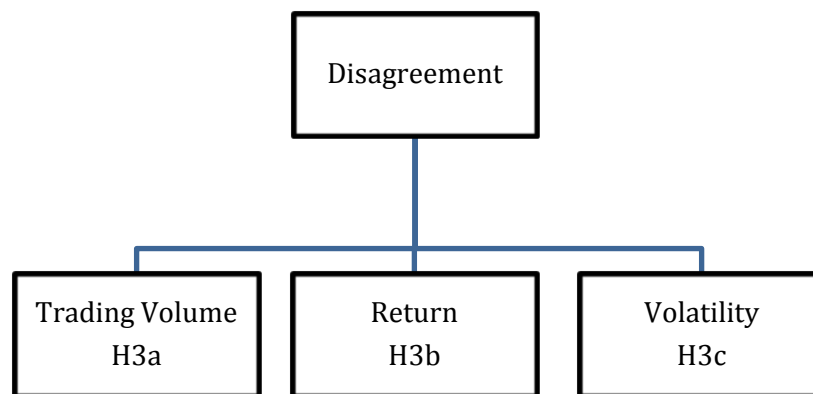


Figure 3.4: The relationship between investor disagreement and stock market variables (trading volume, return and volatility).

▪ Disagreement and Trading Volume

The divergence of opinion among investors in the stock market and its prospective relationship with trading volume has been extensively addressed by scholarly research. Research on the disagreement-volume relation has reached different conclusions about whether disagreement increases or decreases the trading volume. Two distinctive beliefs on whether disagreement affects trading volume were provided by finance theory. The first prospective is found in the “no trade theorem” of Milgrom and Stokey (1982). According to the “no trade theorem”, disagreement can

Chapter Three: Conceptual Framework

actually reduce trading volume because risk-averse investors tend to be aware that the other party to the transaction will only trade if it is favourable to them. Consequently, disagreement can actually result in a decline in trading volume. The no-trade theorem is actually based on the assumption that only a small amount of information is reflected in stock prices (Sprenger et al., 2014).

Another perspective provided by traditional finance suggested that disagreement among investors could potentially impact the volume reaction. This hypothesis implies that investor disagreement may increase trading volume as two market participants allocate different values to an asset (Harris and Raviv, 1993; Karpoff, 1986; Kim and Verrecchia, 1991). For example, Kim and Verrecchia (1991) argue that investors' disagreement before the information is publicly released causes investors to revise their trading strategies and change their beliefs. This differential belief accordingly stimulates trading volume reaction. Bamber et al. (1999) later support this argument and show that investors are consistently updating their beliefs following the information release, which encourages them to trade more, hence increasing trading volume. On the other hand, a recent study by Banerjee and Kremer (2010) contends that even after the information on earnings is publicly available, the on-going effect of disagreement on increasing trading volume is still valid. Their model proposes that trading volume reaction is driven by both divergence and convergence of investor opinions following earnings announcements. In the context of online stock forums, disagreement among online messages can result in an increase in trading volume (Antweiler and Frank, 2004). Given the large number of both noise traders and arbitrageurs participating in stock micro- blogs, one would expect the following hypothesis to hold:

H3a: Disagreement among investors in stock micro-blogging forums has a positive impact on trading volume.

▪ Disagreement and Stock Market Return

The implication of divergence of opinion for the stock market has been addressed by Miller (1977). His theory predicts that, with the existence of short-sell constraints and the disagreement among investors, prices will reflect only the valuation of the most optimistic investors but not the pessimistic ones. In the market,

Chapter Three: Conceptual Framework

if the short-sell constraint binds, investors with high valuations will not short the stock; they either sell the shares or stay out of the market if they agree with the market price. Miller's model suggests that a higher divergence of opinion leads to a higher market price compared to the real value of the stock and lower future returns. Consistent with Miller's model (1977), Diether et al. (2002) find that stocks of high-volatility forecast distributions earn lower future returns when they view the dispersion of analysts' forecasts as a proxy for divergence of opinion.

A considerable number of research papers on the impact of divergence of opinion and earnings announcements have reached different conclusions. While Berkman et al. (2009) show that high disagreement is associated with lower short-window excess returns at earnings announcements, Garfinkel and Sokobin (2006) used unexpected turnover as a proxy for disagreement to show that disagreement at post-earnings announcement is positively related to future return. Although these results seem to lead to different conclusions, they are in fact not definitely contradictory as each study used a different event window of return. Both Berkman et al. (2009) and Garfinkel and Sokobin (2006) apply Miller's theoretical model in predicting return at earnings announcement, which is based on the assumption that the public announcement of information and earnings announcement can help resolve uncertainty and disagreement. However, this assumption may not always hold and this may explain these contrary results. Most recent work by Giannini et al. (2013) observes that disagreement has a negative impact on post-earnings announcement returns.

The forgoing suggests that disagreement among investors has a significant effect on stock market returns. The nature of the posts released in the stock micro-blogging forums along with the related sentiments of investor's opinions allows us to test whether the disagreement by online investors is related to lower stock market returns. This implication of the divergence of opinion and stock market return yields the following hypothesis:

H3b: Investors' disagreement in stock micro-blogs has a negative effect on stock market returns.

▪ Disagreement and Stock Return Volatility

The noise trader model (DSSW, 1990) in behavioural finance suggests that irrational investors often trade on noise not related to fundamentals (Black, 1986), which can affect the volatility of stock returns. The noise traders are acting in concert when buying or selling stocks (Kumar and Lee, 2006), which leads to divergence of the asset's price from the fundamental value. On the other hand, the rational arbitrageurs, who have rational expectations, are always taking an opposite position to that of noise traders (DSSW, 1990). They always try to drive the price back to the fundamental value by selling when noise traders are bullish and pushing prices up and buying when noise traders are bearish and driving prices down. The divergence of opinion among noise traders and arbitrageurs and the unexpected change in noise traders' beliefs will prevent arbitrageurs from taking opposing positions to noise traders, which may create risks that are difficult to diversify. Those risks caused by noise traders are called "noise trader risks". Friedman (1953) and Fama (1965b) both point out that, in the market and in order to bring the price to its fundamental value, rational arbitrageurs are always in disagreement with noise traders and trade against them. The existence of noise trader risk caused by the divergence of opinion between noise traders and arbitrageurs is consistent with Varian's (1985) model showing that investor disagreement by itself is an additional risk factor.

Investor disagreement is sometimes difficult to measure, as investor opinions are not directly observed. For example, the effect of investor disagreement on stock return volatility is indirectly perceived through market information release. Das et al. (2005) argue that disagreement with respect to market information results in extensive debate, which in turn results in the release of new information. While more information is expected to reduce volatility, intuition suggests that there should be a positive relationship between volatility and disagreement in stock micro-blogging forums. Volatility is a reflection of the dispersion of beliefs among market participants (Sprenger et al., 2014). Therefore, this study will derive and test the following hypothesis:

H3c: Disagreement among investors in stock micro-blogging forums has a positive impact on stock market volatility.

Chapter Three: Conceptual Framework

3.5 Chapter Summary

The review of the literature on the predictive ability of online stock forums in predicting stock price movements in the capital market has revealed that stock micro-blogging features have an impact on various financial market indicators in the capital market. Some Tweet features have their own direct relationships with market indicators (i.e. sentiments and agreement) while others (e.g. message volume) show indirect relationships with other market variables. For example, message volume indirectly affects returns and volatility through its effect on trading volume. Therefore, the researcher has developed a conceptual framework, which is based on different theories such as the efficient market hypothesis, random walk and behavioural finance theory. Nine different hypotheses have been developed to establish the conceptual framework to investigate the predictive impact of Stock Micro-blogging in predicting market behavioural movements in the capital market.

After presenting and discussing the conceptual framework of this research, the thesis proceeds in the next chapter to discuss the research design and methodology undertaken and adopted for the empirical field while highlighting the data collection and data analysis methods employed.

CHAPTER 4: RESEARCH METHODOLOGY

4.1 Introduction

The previous chapter proposed a conceptual framework that is aimed at investigating the predictive power of Stock Micro-blogging sentiment in predicting stock market behaviour in capital markets. This chapter aims to provide an explanation of the determination of an appropriate research methodology for guiding the validation of the conceptual framework, hence providing answers to the research inquiries. A philosophical stance is essential in selecting an appropriate methodological approach. Eldabi et al. (2002, p. 64) state: “conducting any type of research should be governed by a well-defined research methodology based on scientific principles”. Therefore, this chapter provides explanations and justifications for the selection of the research philosophy, the type of research approach, data collection methods, and the methodological framework design and data analysis techniques.

This chapter is organised as follows. Section 4.2 provides an overview of the underlying research paradigms. This is followed by an overview discussion of the two research approaches (quantitative and qualitative) in section 4.3. Section 4.4 offers a justification for the most appropriate and preferred philosophical stance that suited this research study and a justification for the selection of the quantitative approach. Section 4.5 addresses the overall research design and explains the most suitable procedures for conducting this study. Section 4.6 explains and justifies the most appropriate methods for data collection. Issues regarding the rationality of the model implemented and data analysis procedures adopted in this study are discussed in section 4.7. Section 4.8 discusses the various components of the framework design while elaborating on the details of each of these components separately. Section 4.9 addresses and explains various financial econometrics models and techniques employed in this study. Finally, section 4.10 offers a brief summary of this chapter.

4.2 Research Paradigm

The scientific research paradigm determines the methodology used in research in order to discover the nature of reality and comprehend its knowledge (Myers,

Chapter Four: Research Methodology

2013). A research paradigm is a set of beliefs, feelings and assumptions about certain aspects of the world and how it should be understood and studied (Collis and Hussey, 2013; Oates, 2006; Guba, 1990), in this instance how to proceed with scientific research (Orlikowski and Baroudi, 1991). Myers (1997) points out that a researcher can have unique beliefs and values and that all research stems from a fundamental assumption of the roots of valid research and what is considered suitable when using research methods. Field research is driven by research paradigms, consisting of three major, mutually connected beliefs about ontology, epistemology and methodology, as shown in Table 4.1. A discussion of the philosophical arguments directed at the various research paradigms is not the main concern of this section; rather, it aims to explore the context of research and define the epistemological approach in this study. However, it will discuss and stress the main aspects related to the different philosophical approaches in the fields of economics and finance. Pozzebon (2004, p. 277) states that the research paradigm helps researchers find their true course and gives them a stepping-stone towards the value of what they wish to accomplish.

Table 4.1: Ontology, Epistemology and Methodology: Differences between positivist and interpretive research paradigms

Basic Beliefs	Research Paradigms	
	Positivist	Interpretive
Ontology (What is the nature of “being”/reality?)	<ul style="list-style-type: none"> - The world is external - Only single objective reality exists. 	<ul style="list-style-type: none"> - Multiple realities exist.
Epistemology (How is reality captured? The relationship between the researcher and reality)	<ul style="list-style-type: none"> - Only “facts” derived from the scientific method make legitimate knowledge - Researcher is independent and does not affect the research outcomes 	<ul style="list-style-type: none"> - The researcher is involved in what is being researched.
Methodology (How should the researcher go about social reality?)	<ul style="list-style-type: none"> - Survey Questionnaire - Simulation - Experiment - Cross-Sectional - Correlational - Formalised statistical and mathematical methods 	<ul style="list-style-type: none"> - Action Research - Case-Study

Source: Compiled after Bailey (2007) and Vaishnavi and Kuechler (2008)

Chapter Four: Research Methodology

There are two main approaches to selecting research methods, namely positivist (Hussey and Hussey, 1997), and interpretivist (Mingers, 2001). The positivist approach, commonly known as the scientific approach, is normally based on quantitative methods, while the interpretivist approach normally encompasses qualitative methods. Both philosophical approaches, depending on the circumstances, may have positive or negative effects on certain research projects, although the main concern still remains (Bryman, 2012). The next section looks at these approaches more closely and provides reasoning for selecting one or the other research philosophy.

4.2.1 Positivist Paradigm

In the positivist paradigm, inquiry is considered to be value-free, and researchers are unconcerned and indifferent or, rather, neutral and unbiased (Collis and Hussey, 2013; Easterby-Smith et al., 2012). As explained in Collis and Hussey (2013), positivism which is based on the approach used in natural science, has its roots in the philosophy that is known as realism and it is based on the belief that reality is independent of humans where the aim is to discover theories based on empirical research via observations and experiments. As Chen and Hirschheim (2004, p. 201) put it, “positivists believe that reality exists objectively and independently from human experiences.” From a philosophical standpoint, the paradigm in positivism is deductive, starting with developing hypotheses from theory and followed by collecting data. For the purpose of finding the general principles or laws that govern the natural and social world and boost the predictive understanding of the investigated phenomenon, positivist research focuses on the empirical testability of theories (Myers, 2013; Orlikowski and Baroudi, 1991). The pivotal concern of positivists is how to utilise some random sampling techniques, measure outcomes and design causal models with a definite, substantial, predictive factor (Myers and Avison, 2002). To Orlikowski and Baroudi (1991), positivist research must entail evidence of formal propositions, quantifiable measures of variables, hypothesis testing and arriving at conclusions regarding a phenomenon from the very sample to the given population. Additionally, from a mathematical point of view, positivist researchers assume the possibility of generalising and modelling the observed phenomena (Oates, 2006).

Chapter Four: Research Methodology

According to purist quantitative researchers (Maxwell and Delaney, 2004; Popper, 1959; Schrag, 1992, as cited in Johnson and Onwuegbuzie, 2004), social surveys should be undertaken with a similar approach to that taken by physical scientists when processing physical phenomena. Furthermore, they consider that by maintaining neutrality and objectivity from their subjects (research participants), research processes are significantly improved. This translates into the science of generalisations, which requires that social science inquiry be objective, and it is perhaps the only way to gauge the validity and reliability of the real causes of social scientific outcomes (Nagel, 1989). The positivistic school of thought opines that researchers must not become emotionally involved and must remain critical of the objects and participants of the study. As a result, positivists try to embrace the neutral side, using a formal writing style along with the passive voice and technical terminology (Tashakkori and Teddlie, 1998). Natural sciences and the study of natural phenomena paved the way for the creation of the research methods that are now used by positivists. Numerics, laboratory experiments, simulations, mathematical modelling and econometrics are most commonly used in the field of economics and finance (Myers, 2013; Neuman, 2005; Myers and Avison, 2002). The quantitative research approach is based on deductive reasoning. A postulate is set a priori, and data are gathered to test the validity of the hypothesis.

4.2.2 Interpretive Paradigm

The interpretive paradigm is based on an ontology in which reality is subjective, a social product constructed and interpreted by humans as social actors according to their beliefs and value systems (Andrade, 2009; Saunders et al., 2011). Interpretivists eschew any research in which one stands by as a neutral observer and emphasise human interpretation and comprehension in the light of valid knowledge (Gray, 2013, Saunders et al., 2011). Unlike positivist research, an interpretive study aims not to prove or disprove a hypothesis but, rather, “to identify, explore and explain how all the factors in a particular social setting are related and inter-dependent” (Oates, 2006, p. 292).

Qualitative purists – constructivists and interpretivists – adopt the “the superiority of constructivism, idealism, relativism, humanism, hermeneutics and, sometimes, postmodernism” (Guba and Lincoln, 1989; Smith, 1983, 1984, cited in

Chapter Four: Research Methodology

Johnson and Onwuegbuzie, 2004, p.14). For qualitative purists, time- and context-free generalisations are not considered acceptable because of the existence of multiple constructed realities. The debate extends further to the point where it is hard to make the difference between causes and effects that stem from specific generalisations. Guba (1990) points out that knower and known are inseparable, making the perceived reality subjective rather than objective. Qualitative purists use a detailed description containing rich information, while quantitative purists maintain a more formal style of writing.

According to the interpretivist school of thought, subjective interpretation and intervention are required in order to understand how reality works (Davison, 1998). According to interpretivists, reality cannot be defined objectively; rather, it is expressed socially (Hussey and Hussey, 1997). What makes this notion sustainable is that the possibility of understanding people's perceptions is far greater with regard to those activities that they do socially. Interpretivists tackle the significance of qualitative data with great care in the development of knowledge (Kaplan and Maxwell, 2005). As such, the methods used in qualitative research that were developed in the social sciences served researchers studying social and cultural phenomena. Qualitative research is based on induction. The process involves gathering data, examining them, and finally constructing theories from all the evidence. Table 4.1 provides a summary of the main differences between these two approaches.

4.3 Research Approach

Research can also comprise both quantitative and qualitative approaches. The two sections below provide further details of the two approaches and examine why one is more useful than the other for this particular study.

4.3.1 Quantitative

Quantitative research roots go as far as the natural sciences in the study of natural phenomena (Saunders et al., 2011). This approach to research focuses on measurements to describe objects and relationships (Sarantakos, 2005). Moreover, researchers who use quantitative methods pay little regard to the context; rather, they

Chapter Four: Research Methodology

emphasise large numbers regardless of the context of the data, looking for statistical significance (David and Sutton, 2004; Neuman, 2005). Examples of quantitative research methods are survey questionnaires, laboratory experiments, simulations, mathematical modelling and econometrics (Myers, 2013; Neuman, 2005; Myers and Avison, 2002).

4.3.2 Qualitative

Compared to quantitative research, qualitative research is based on words or pictures rather than numbers (Johnson and Harris, 2002; Miles and Huberman, 1994). For the purpose of understanding human behaviour, qualitative research focuses on illustrating how data are collected. The qualitative approach is recognised as a phenomenological, subjective or non-positivistic approach. If the issue or problem is unknown, it is practicable to use qualitative methods. Qualitative researchers, then, unravel the meaning of the experiences people find when tackling certain issues (Sarantakos, 2005). Researchers who use qualitative methods tend to use small samples and study them extensively in their original form (Berg, 2014). By conducting face-to-face interviews, researchers are able to study their subjects' behaviour and reactions within the context itself (Creswell, 2012). Gibbs (2002, p. 3) argues that the qualitative method "involves a commitment to viewing events, actions, norms, values etc. from the perspective of those understudied". Likewise, Creswell (2012) points out that qualitative research provides the means to gain an understanding of the social phenomenon and its meaning. Human behaviours can be profoundly understood using this method, by taking into consideration people's values, interpretive schemes and belief systems (Cavana et al., 2001). Individuals are placed in this philosophy for the purpose of clarifying how and why a phenomenon occurs (Sharif, 2004), trying, at the same time, to illustrate the actuality by giving accurate descriptions as seen by the participants in order to provide meaning to the fundamental human actions (Sarantakos, 2005). This method is an alternative for collecting data by the positivistic approach, which is about interpretive research (Neumann, 2005).

Chapter Four: Research Methodology

4.4 Justifications for the Adoption of the Research Approach

The approach adopted by this study is positivist (quantitative) and, having defined the two research paradigms, the focus is on investigating how well stock micro-blogging sentiments can determine stock price movements in the capital market. The reasons why positivism paradigm was chosen are provided below.

Firstly, the study seeks to evaluate what is missing by utilising certain theories and models in the fields of economics and finance, and to examine and, ultimately, test given hypotheses and quantifiable measures of variables. To determine these hypotheses, theories are tested through empirical investigations using the most appropriate deductive methods. Philosophically, the positivist paradigm uses deductive methods by identifying what is missing from the literature and creating hypotheses from existing theory; these are then analysed and tested. This research aims to produce quantitative evidence, with much less regard to interpretivist epistemology.

Secondly, the type of data collected (secondary data), the nature of the data collected (StockTwits data and historical financial data) and the purpose of the study (measuring and investigating the relationship between stock micro-blogging features and financial market indicators) inherently propose a quantitative positivism paradigm as the most suitable approach for this study. The researcher is able to choose the most appropriate techniques related to the data collection and analysis methods, thus maintaining neutrality throughout the research (Hussey and Hussey, 1997).

Thirdly, explanations are provided for the dependent and independent variables for the purpose of obtaining generalisable findings. The goal of this research is to investigate investors' sentiments from an online Stock forum for their capacity to change stock prices in the capital market, investigating whether stock prices can be predicted at all.

Fourthly, the nature of this research study is empirical, using data mining techniques and various econometric models. This research uses the Dow Jones Industrial Average (DJIA) Index while the StockTwits data and financial market information are collected from about 30 companies in the DJIA index over a one-year period from April 3rd 2012 to April 5th 2013. Different kinds of text mining techniques (statistical and analytical) were used to analyse the data of StockTwits

Chapter Four: Research Methodology

posts and their sentiment measures (e.g. bullishness, message volume and agreement), which are relevant for financial indicators such as returns, volatility and trading volume. Since this research mainly emphasises the use of numbers in an objective fashion and statistical methods are used for the analysis, the (positivist) quantitative approach is considered suitable for this study.

Finally, ontology, ultimately, suggests the realist position that requires social facts, whereas epistemology embodies human facts and causes. Two aspects are highlighted in this research, namely the realism of the context and the use of quantitative research methods such as facts and causes of social phenomena - stock prices prediction in financial markets. Its starting point is that the social world comprises relevant empirical artefacts, which are identifiable, can be studied and, therefore, can be measured using natural sciences approaches. Burrell and Morgan (1979) stated that quantitative research focuses on demystifying events in the social world by analysing regularities and causal relationships. According to Gilbert (2008), the goal is to build a concrete path to the collection of “facts” about society so that explanations might be given through statistical evidence of the way the social world remains in its orbit.

4.5 Research Design

The research design is defined as the “science (and art) of planning procedures for conducting studies to get the most valid findings” (Vogt, 1993, p. 196). A research design is the plan devised by a researcher to investigate the topic and help answer the research questions (Cooper and Schindler, 2001). It is a cohesive and coherent procedure for conducting studies to collect, analyse and interpret data. According to Yin (2013, p.13), the appropriate methodological design has to fit with “(1) the research problem, (2) the extent of control the researcher has over actual behavioural events and (3) the time-focus of the phenomena observed, i.e. contemporary or historical”. The primary purpose of the research design is therefore to enable a researcher to determine the research boundaries that appropriately address the research problem, the kind of investigation and the analytical procedures that need to be undertaken, the unit of analysis and other related research issues. Hussey and Hussey (1997) argue that the precise selection of the research design is one of the most important success factors of any research study.

Chapter Four: Research Methodology

In this study, three types of research designs identified from the literature, namely (1) exploratory, (2) descriptive, and (3) explanatory design, will be employed. In the early stages of the research study, the exploratory design was adopted to identify the research problem, to critically review the literature (Churchill, 1999), and to create the theoretical framework to generate the hypotheses based on previous empirical studies, as reported in chapter 3. At this point, the research problem was defined and the research aim was clearly identified; i.e. this research study focuses on understanding the predictive ability of stock micro-blogging sentiments in predicting stock market behaviour. The descriptive design of the research is then used to accomplish the text-mining task to extract sentiments from the collected StockTwits. Different textual analysis techniques and machine learning algorithms are used to analyse StockTwits posts. A descriptive design was also used to perform the predictive analysis approach using different statistical techniques to understand the predictive relationship between the market features (e.g. market return, volatility and trading volume) with corresponding stock micro-blogging sentiments (e.g. bullishness, message volume and level of agreement). Sometimes, however, the descriptive research design may not fully explain the association between the studied variables (Zikmund, 2000). In order to fully explain the relationship and association between variables, the explanatory research design is highly important at this stage.

4.6 Data Collection

Data collection is an essential component of research design that is used to collect empirical research data. It is a technique that shows how researchers obtain the information they need to conduct their study.

4.6.1 Data Generation Sources

Data sources are generally classified into two broad categories: primary data and secondary data. Primary data are collected and observed directly from first-hand experience and are generated by researchers for a specific research purpose. In contrast, secondary data are data that already exist, are available and have not been collected for a specific research purpose (Johnson and Turner, 2003; Sorensen et al., 1996; Lehmann, 1989). In secondary sources, the information is offered in either written or electronic form. There are several types of secondary data that researchers

Chapter Four: Research Methodology

may use, such as personal documents, official documents, physical data and archived research data. Regardless of the data categories (primary or secondary) being used, validity is an important issue that researchers must be aware of in the data collection stage. The term ‘validity’ simply refers to the awareness that one is conducting high-quality research (Johnson and Turner, 2003). As Johnson and Christensen (2000) state, research is said to be valid if it is plausible, credible, trustworthy and defensible. Lincoln and Guba (1985), however, used ‘trustworthiness’ to refer to high-quality research.

- **Justification for the adoption of the secondary data source for data collection**

This research study proposes to use secondary data as a method of data collection, specifically “documentary-based secondary data” and archived research data. Data are collected from two secondary data sources: the online text from a stock micro-blogging forum (StockTwits messages), which can be classified as “documentary-based secondary data”, and the financial market data (daily stock prices on the Dow Jones Industrial Average (DJIA index) from the US market where all companies are quoted in the National Association of Securities Dealers Automated Quotations (NASDAQ) and the New York Stock Exchange (NYSE), which can be classified as archived research data. There are a number of reasons for valuing the information obtained from utilising secondary data (Houston, 2004). First, secondary data are a representation of “real” decisions that have been made by “real” decision-makers in different aspects of “real” life (Winer, 1999). Therefore, information gathered from an investment forum such as StockTwits, where people are sharing real-time investment ideas to help investors make profitable investment decisions is considered real-time data produced by “real” investors about “real” investment life. In StockTwits, people post very frequently and close to the occurrence of events in real time, thus producing up-to-date information posts. This information is collected in a less obtrusive manner compared to a survey data collection where people may be unwilling to share their ideas and may refuse to complete the survey.

Second, micro-blogging services such as StockTwits are a more effective and efficient source of data for use in this research. It provides readily available data at low cost, allowing a faster and less expensive extraction of features and indicators for

Chapter Four: Research Methodology

the study (Oliveira et al., 2013). Moreover, the short message posts in StockTwits make them a source of data that are very easy to handle, with less noise embedded in them.

Third, the reality-based nature of the secondary data may provide an indication that secondary data reduce the likelihood of self-reporting biases that may be present in the other forms of primary data collection (Houston, 2004; Johnson and Turner, 2003). “[S]elf reports can be influenced by a variety of factors, including self-presentational concerns or what has been termed ‘self-deception’” (Tomarken, 1995, p. 388). For example, the financial market data obtained from Bloomberg (professional financial terminal) that are used in this study are provided to the market in accordance with the Security Exchange Commission (SEC), thus preventing the possibility of any reporting biases by researchers or other participants. Fourth, secondary data are a wonderful source of data for answering exploratory questions (Windle, 2010). Since our research questions are of an exploratory type, using StockTwits posts and other financial market data as a source of secondary data enables a more precise interpretation of the data. It also allows for greater control over the factors impacting the validity and reliability of the data being analysed. Finally, the type of data used in this research study relied on financial market information, which is historical data obtained from a professional financial database in the form of historical records, which is regarded as a form of secondary data.

Table 4.2: Secondary data collection method: strengths and limitations	
Strengths	<ul style="list-style-type: none">• Cost-effective, efficient, and convenient method for collecting data by researchers.• Widely exist and available for research purposes.• Exact and contain precise details of information required for a particular research study.• Inexpensive and less time-consuming• Reduce likelihood of bias
Limitations	<ul style="list-style-type: none">• Researcher control over the data is possibly low.• Sometimes difficult to validate the data being collected.• Ethical considerations for the use of secondary data must be met.

Source: Adopted from Houston (2004) and Sorensen et al (1996)

There are other broad advantages of using secondary data sources. One of these advantages is that they already exist and are available. Therefore, the time spent on research using secondary data sources is less than the time spent on conducting

Chapter Four: Research Methodology

studies that use primary sources of data (Houston, 2004; Sorensen et al., 1996). Furthermore, secondary data are considered inexpensive and a less costly method of data collection compared to other types of primary data collection (Johnson and Turner, 2003). Table 4.2 highlights the key strengths and weakness of the secondary data collection method that was employed in this study.

4.6.2 Instruments

A number of useful tools and instruments will be used to collect data for this research study. The following sections will discuss each of these instruments in more detail.

- **StockTwits**

StockTwits is an online financial communication platform where a group of highly talented and motivated people come together to share and receive investment ideas about the financial market. StockTwits is one of the leading communications platforms for the financial and investment community, where real-time streams of investment ideas flow among people on the platform. With StockTwits, investors and traders are better informed than ever before, which may help them manage their investment portfolios better and make better investing decisions. StockTwits offers a specialised financial atmosphere where investors, institutions, stock brokers, economic consultants and market analysts come together to donate, find and track ideas about stocks, markets, trends and more. As Lindzon et al. (2011) state, StockTwits can help investors to get the ideas they want from the people of interest to them and about the stocks that matter to them most (Anonymous, 2011). In addition, StockTwits' integration with leading social networks and financial sites, including Twitter, Facebook, LinkedIn, Yahoo Finance, CNN Money and Reuters (“About Stocktwits”, 2013), allows investors and users to broaden their social interactions and view a large number of opinions of other people on other platforms.

StockTwits is a financial blogging forum founded in 2008; it created the \$TICKER tag to help organise the stream of information around stocks and the market across the Web and social media platforms. It is a financial communication platform that allows more than 300,000 investors, financial professionals, market

Chapter Four: Research Methodology

analysts and public companies to contribute and share investment ideas and opinions about individual stocks and the markets (“About StockTwits”, 2013). It allows users to monitor the activities of traders and investors through market conversations on the platform, which may alter their investment decisions on the stocks of interest.

▪ Financial Market Data

The stock market data issued at daily intervals are obtained and downloaded from Bloomberg for the Dow Jones Industrial Average (Dow 30) covering the period from April 3rd 2012 to April 5th 2013 (Bloomberg, 2013). Bloomberg is the most powerful DataStream providing real-time financial data for financial professionals and researchers alike. This research study will focus on the DJIA index to adequately reflect the US stock market. The DJIA is a price-weighted average of 30 large ‘blue-chip’ stocks traded on the New York Stock Exchange (NYSE) and the National Association of Securities Dealers Automated Quotations (NASDAQ). Regardless of the limitations in the composition and structure of the index, it is nevertheless the most widely followed and reported stock index (Lee, et al., 2002). The DJIA is particularly well suited to this study because it constitutes the large capitalisation industrial companies of the US equity market. The 30 stocks making up the index comprise about 25 per cent of the market value of the entire NYSE (Lakonishok and Smidt, 1988). Therefore, a focus on large and highly traded firms will probably reduce the problems associated with non-concurrent trading (Rudd, 1979). This in fact makes the DJIA a reasonably valuable index for representing short-term market movements. In addition, since the companies that make up the DJIA are actively-traded companies, their stocks generate a greater ‘buzz’ on social media networks. Therefore, these stocks are heavily discussed in StockTwits and are the subject of a very high volume of tweets. Moreover, the DJIA index constitute companies from different sectors in US market (i.e, Financial, Oil and Gas, Consumer Goods, Health Care,..etc.) However, this research study may be valid for any companies/indices that generate a low volume of tweets. The price data is obtained for each company constituting the Dow Jones index on daily bases over the study sample period.

4.6.3 Statistical Packages

▪ Text Mining (TM)

The `tm` is a software package provided in R⁸ that offers functionality of classical applications (Weiss et al., 2010) for managing text documents, such as document clustering (Zhao and Karypis, 2005) and document classification (Sebastiani, 2002). Natural language processing techniques are then applied to transform the text according to term frequencies into highly structured representations. `Tm` is an open source package, which provides a framework enabling researchers and practitioners to organise, transform and analyse textual data. This package is publicly available to researchers via extension packages (e.g `kernelab` and `lsa`) or via interfaces with established open-source toolkits from the data/text mining field, such as Weka or OpenNLP from the natural language processing community (Feinerer et al., 2008). With the `tm` package being integrated into R, a highly sophisticated model has been built for text-mining purposes, with leading statistical computing methods made available to researchers.

In this research study, online text from an online stock forum called “StockTwits” has been collected to perform the text-mining task. In fact, the online text is a rather unstructured collection of words, which makes textual analysis a challenging process. Since the texts obtained from StockTwits are all in the English language, the `tm` package is an ideal tool for text mining to build text classification models for the collected ‘tweets’ from StockTwits. `Tm` is an application that makes use of the functionality provided by the Weka toolkit.

▪ WEKA Toolkit

Weka is machine learning software written in Java that implements many state-of-the-art machine-learning and data mining algorithms (Witten et al., 1999). There are a number of reasons why Weka is being used in this research study. First, since three different machine-learning algorithms are used and compared to perform the feature selection task, Weka is the most appropriate software to use as it provides tools for analysing the resulting classifiers and allows a performance evaluation comparison (Witten et al., 1999). Second, Weka also provides tools for pre-processing

⁸ R is one of the leading computing environments for statistical applications and graphics (the R project website, <http://www.r-project.org/>).

Chapter Four: Research Methodology

data routines including feature selection, which is the data-mining tool employed in this research study. Third, Weka offers different measures for evaluating relative performances of several learning algorithms and verifying the robustness of models (e.g. cross-validation). Fourth, Weka can navigate the data automatically from the generated source through its online documentation (Witten et al., 1999).

4.7 Data Analysis

Data analysis is the process of systematically applying analytical and/or logical techniques to search, review, illustrate and evaluate the data for the purpose of gaining understanding and finding useful meaning (Boeije, 2009). According to Shamoo and Resnik (2003), different analytic procedures provide a technique of drawing inductive inferences from data while differentiating the relevant signal from the noise present in the data. Data analysis is one of most important steps that must be completed when conducting a research study. It is a phase of research that follows the collection of data, including classifying, coding and tabulating data needed to perform quantitative or qualitative analysis based on the research design applied and the appropriateness of the data. Therefore, deciding how to analyse the data prior to data analysis is a critical decision to be made by researchers to avoid data being collected in an improper format and to prevent inaccurate findings from the data (Cooper and Schindler, 2001). There is no specific formula or standard technique for analysing quantitative data. In fact, there are a variety of techniques available to perform this task; research methodology books and articles offer various general analytical approaches to quantitative data, including questionnaires, text analysis, data mining, data visualisations and statistical analysis (Lemon et al., 2010; Bryman and Cramer, 2001) In this research study, data mining, text analysis, and computational/statistical approaches to financial analysis (both descriptive and inferential statistics) will all be applied. The later sections of this chapter will explain and discuss the various text mining techniques and econometric analyses, which were found to be the most appropriate methods for this research.

Chapter Four: Research Methodology

4.7.1 Rationale of Modelling

This thesis emphasises Text Mining (TM) as the most suitable strategy for this particular research. This study follows the view of Bernard and Ryan (1998) that the most appropriate research strategy for conducting social science empirical research involving the systematic analysis of financial text in the sociological tradition is textual analysis. The sociological tradition is a type of text mining that treats text as a window into human practice (Bernard and Ryan, 1998). Mckee (2003) defines textual analysis as a data-gathering process and methodology that enables researchers to gain an understanding of the ways in which individuals of a given culture make sense of who they are and how they really fit into the world in which they live. Textual analysis has been defined as “a research method for the subjective interpretation of the content of text data through the systematic classification process of coding and identifying themes or patterns” (Hsieh and Shannon, 2005, p.1278). Research employing textual analysis often attempts to manage and handle massive amounts of information to extract meaning and knowledge from a given text through the application of computer-based methods and techniques (Fan et al. 2006). The overall goal, essentially, is to convert text into data for further analysis via the application of data mining techniques and machine-learning algorithms. The nature of the data collected (StockTwits posts) as well as the purpose of the data analysis (to extract sentiments from online financial text) inherently proposes the need for text mining to conduct this study.

The rationale of the model in this thesis is that the models are trained from a corpus of manually labelled data to test the computational model instead of using a sentiment lexicon, such as the SentiWordNet. The SentiWordNet is a lexical resource used in most opinion-mining tasks to assign three sentiment scores to a text: positivity, negativity or objectivity (sentiwordnet.isti.cnr.it). It has been widely used in multiple research papers to perform three tasks: to determine the subjective/objective polarity of a text, to determine the positive/negative polarity of a text, or to determine the strength of a text’s positive/negative polarity. The main reason why existing lexicons are not used in this thesis is that this research is based on extracting sentiments from financial texts, as it has been decided to classify text as buy, sell or hold, and not merely as positive or negative. The vast majority of research papers in the sentiment analysis field mainly focuses on areas including emotional states (Kramer, 2010),

Chapter Four: Research Methodology

product reviews (Turney, 2002) and movie reviews (Pang et al., 2002), in which cases SentiWordNet is deemed a suitable lexicon. However, financial researchers have shown that dictionaries developed from other disciplines may not be effective in financial texts and may result in a misclassification of common words (Loughran and McDonald, 2011). For example, for the StockTwit “*Short \$MSFT @ 29.18*”, if the SentiWordNet lexicon is used, the sentiment will probably be assigned as objective or neutral, while in the context of finance the word *short* is a vibrant sign signifying that the participant expects the Microsoft Corporation (\$MSFT) stock to fall.

The following sub sections discuss the methodological issues associated with text analysis and tweet sentiment hand-labelling. These two issues are sampling and coding, which are the heart of textual analysis.

4.7.2 Sampling

There are two elements related to the sampling issue: the identification of the corpus of the text and the identification of the unit of analysis within the texts (Bernard and Ryan, 1998).

- **Identifications of the Corpus of the Text**

Most researchers argue that when the units of data being analysed run into hundreds or even thousands, then a representative sample of the data must be made (Cohen, 1990; Gilly, 1988). Since thousands of StockTwits posts (289,024 valid postings) have been collected to conduct the current study, the models are trained from a corpus of manually labelled posts (2,892 tweets) that have been randomly selected from all companies included in the analysis from different time spans. In random sampling, every element in the population has an equal chance of being part of the sample. There are two techniques for performing random sampling: truly random sampling or systematic sampling (Tashakkori and Teddlie, 1998; Teddlie and Yu, 2007). In truly random sampling the researcher makes the sample selection without taking into consideration any factors that may cause selection biases. On the other hand, in systematic random sampling the sample is selected according to certain predefined rules; for example, a sample can be selected based on a certain percentage of the total population. The main advantage to be gained by using the systematic

Chapter Four: Research Methodology

random sampling technique is that a more representative sample of the target population will be achieved when the population is large. The systematic random sampling method was therefore used in this study where the corpus of manually labelled posts is picked according to two simple rules as follows:

- 1- The sample of tweet posts selected should comprise an equal number of posts from all thirty companies in the DJIA index (Approximately 96 tweet posts per company).
- 2- The sample should be collected from different time spans to avoid any bias in the selection process (i.e. tweets selected from different months of the year, from different days of the week and from different times of the day).

- **Identifications of the Unit of Analysis**

Once the sample of the text is selected, the next step is to specify the unit of analysis (Krippendorf, 2012). The unit of analysis refers to the basic unit of text to be classified and analysed during textual analysis (Zhang and Wildemuth, 2009). Specifying the unit of analysis is critically important for the text mining task. In this model where StockTwits postings will be analysed, the entire tweet's message text is considered the most appropriate unit of analysis for this research.

- **Coding**

Coding can be defined as the process of assigning a unique code to data for the purpose of classification or identifications. As Bernard and Ryan (1998) point out, codes can be used either as indexing to tag text in a corpus or as a measurement device to value text, such as the frequency, amount, or presence/absence of information. At this stage of analysis, the researcher should develop categories and a coding scheme that can be derived from three sources: the data, related literature and theories. Inductive and deductive coding can both be used to develop categories and a coding scheme (Zhang and Wildemuth, 2009). The former coding is particularly appropriate for studies intended to develop theory, whereas the deductive coding is appropriate in confirmatory research to verify an existing theory (Bernard and Ryan, 1998). In this thesis, deductive coding is the most appropriate type of coding because the codes already exist based on previous research and they make it possible to confirm the theoretical framework on whether stock price behaviour can be predicted,

Chapter Four: Research Methodology

as well as to investigate the role of noise traders in the capital market (further details were provided in chapter three). The following subsections discuss the development of categories and the coding scheme that will be applied for the current study.

- **Categories**

In this thesis, manual classifications of a total of 2,892 tweets on all of the 30 stocks are labelled in three distinct categories {buy, hold or sell}, based on a predefined dictionary, the Harvard IV dictionary. The Harvard Psychosocial Dictionary, called Harvard IV, provided by the best-known system, on the General Inquirer (GI) website, contains about 4,000 emotional words classified as either positive or negative (more details will be provided in section 4.8 of this chapter).

- **Coding Scheme**

To ensure the consistency of manual coding when multiple coders are involved, the researcher needs to develop a coding book/scheme. A coding scheme can be defined as the coding rules that coders must follow in the task of assigning labels to the texts to enhance the coder's understanding of the categories, which may lead to greater consistency in the coding process (Zhang and Wildemuth, 2009). The codebook usually contains the category names with clear definitions of each of these categories as well as a clear statement of the rules for assigning codes (Zhang and Wildemuth, 2009; Weper, 1990). Table 4.3 presents the coding scheme used by the researcher and an independent coder for the manual classification of 2,892 StockTwits posts, which will then be used as a training set to test the model's accuracy.

- **Inter-coding Agreement Methods**

When using hand-coded data in which data are labelled in categories (in this research, text is classified into three distinct classes: sell, buy or hold), it is very important to show that such coding is reliable. Artstein and Poesio (2008) argue that there are different ways to test the reliability of the coding, depending on how agreement is tested. If the coding process is performed by the same coder, the reliability will be tested by intra-coder agreement (stability): the extent to which the

Chapter Four: Research Methodology

coding process yields the same results by the same coder when repeated over a distance of time. On the other hand, if two independent coders perform the coding then, the reliability will be tested by reproducibility: the extent to which different coders working independently achieve the same coding. In this research study, the manual classification of StockTwits messages (2,892 tweets) performed by the researcher (the primary judge). In line with most studies based on text classification methods using manual training sets (e.g. Antweiler and Frank, 2004b; Sprenger et al., 2014), a second judge worked independently to perform a manual classification of the same training sets using the coding scheme shown in Table 4.3 to achieve greater reliability and consensus regarding their classification.

In this thesis, the Kappa statistic is used to test the inter-coder agreement reliability. Cohen's Kappa, or K , is a popular statistical measure of the degree of agreement between two independent coders for categorical items (Cohen, 1960). The Kappa statistics will be calculated based on the following formula:

$$\hat{K} = \frac{P_0 - P_e}{1 - P_e}$$

where P_0 is the observed proportional agreement between categorical variables and P_e is the expected agreement between categorical variables by chance. Generally, the K takes a value between 0 and 1 ($0 \leq K \leq 1$), although negative values do sometimes occur. The most critical question regarding the reliability test is this: how much inter-rater agreement is sufficient? There is no cut-off point to determine the agreement rate as the standard is still evolving, but most researchers (e.g. Krippendorff, 2012: 147–148) advocate that agreement of at least 0.70 can be considered good. Further details about the inter-coder agreement methods and results will be provided in Appendix I.

Table 4.3: The Coding Scheme for Manually-labelled Tweets

<p>There are a number of general rules applied in labelling the StockTwits data that are used as input data (training set) in the text-processing model; these are listed as follows:</p>
<ul style="list-style-type: none">(i) If the tweet post contains external links to long articles or numerical charts about the stocks, it is generally marked as neutral. The content of the article and the information revealed by the chart are not taken into account.(ii) Buy, hold or sell labels are only given when the sentiment can be explicitly speculated from the tweet.(iii) Tweets with question marks are generally marked as neutral.(iv) Simple summarisations of the stock performance by the end of the day are not taken into consideration.(v) If the user reports a loss in a subjective way instead of reporting numbers, it is fair to assume that the user has a negative feeling towards the stock and vice versa.(vi) If a tweet post contains company names (Apple, Google, Microsoft) or any other neutral words (such as day, report, look, watch...etc.), it is generally marked as a hold message.(vii) All positive words/emotions in a tweet message indicate linguistic bullishness (e.g. strong, high, happy, earn ...etc.) and will therefore be marked as buy messages.(viii) Sell messages contain corresponding bearish words (e.g. loss, weak, low, fall, decline, down, etc.); therefore, all negative words/emotions in a tweet message indicate linguistic bearishness and are commonly marked as a sell signal.(ix) Normally tweet posts containing a balance of positive and negative words will be classified as hold messages.(x) A tweet post containing a mixture of positive and negative emotional words will be assigned to the correct class based on the probability value assigned to each class where the message will be assigned to the class of high probability. For example, if a tweet message contains 65% positive words, 20% negative and 15% neutral words, the message will be classified as a buy message since positive words are more likely to be associated with the buy signal.

4.8 Framework Design

The previous chapter describes the methods and algorithms that have been chosen to build the framework design for data analysis adopted for this thesis. In order to accomplish the objectives explained in chapter one, a prediction framework is developed and illustrated in Figure 4.1. As shown in Figure 4.1, the framework design is composed of six major components: Data Description and Pre-Processing Framework, Feature Selection and Construction Framework, Text Processing Model, Performance Evaluation, Training and Testing, and Statistical Summaries. These framework components are represented in dashed boxes identified with relative component names. Each component of the framework consists of different procedures

Chapter Four: Research Methodology

that are vital for performing the whole function of the relative component. The Data Description and Pre-Processing Framework is the first component, appearing at the top of the figure; it is responsible for data acquisition from various sources as well as pre-processing and filtering procedures to avoid irrelevancy in the data being collected. At this stage and after the text customisation has been performed, the manual operation of sample tweet messages is performed to manually classify tweets into three distinct classes - sell, buy or hold - using the Harvard IV dictionary; these are then used as a training set in the text processing model and feature construction stages. The second component (Feature Selection and Construction Framework) represents the implementation of two approaches of feature selection (Filter and Wrapper) to extract the most relevant features from the datasets to build a features construction model. The construction model of relevant features (reduced features) is then used as input variables to the third component (Text Processing Model) where three machine learning models (Naive Bayes classifier, Decision Tree and Support Vector Machine) are employed to process the text and detect relative sentiments. As three different models are used for text processing and each model has its own bias and assumptions that will lead to different accuracy levels, it is important to evaluate the model performance, which is the task handled in the fourth component of the design (Performance Evaluation). As it can be seen from the design, this component involves a number of methods that are used to validate the models and evaluate their performances in order to identify the best model with the highest accuracy level. The fifth component (Training and Testing) is an important step in the data mining task. It is used to examine whether the classifier has the qualifications and ability to predict any new instance emerging from the environment. The final component (Statistical Summary) involves the application of various statistical tools to statistically measure the relationship between StockTwits features, such as bullishness, message volume and level of agreement, and financial market indicators such as trends, trading volume, return and volatility.

Chapter Four: Research Methodology

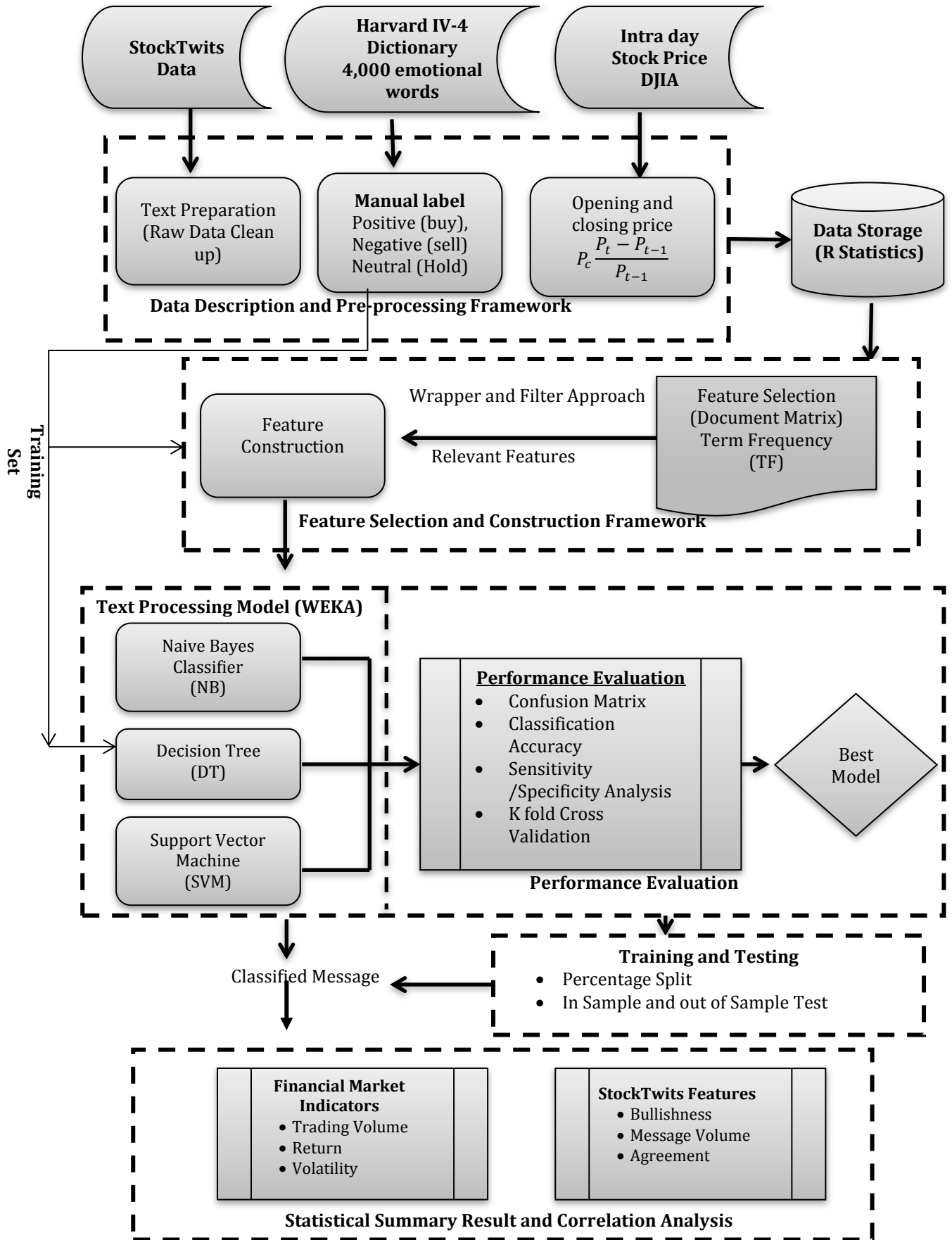


Figure 4.1: Framework Design

Chapter Four: Research Methodology

The rest of the chapter explains in more detail the function of each framework component along with the role of each individual process depicted within it.

4.8.1 Data Description and Pre-Processing Framework

The Data Description and Pre-Processing Framework is the first component of the data analysis design; it is responsible for describing the nature of the data that have been acquired from different sources, as shown in Figure 4.1 Two main sources have been used to acquire the data for this research study: StockTwits data and Financial Market data. General Inquirer's Harvard IV-4 Dictionary also appears as data input in the first component of the design as it can then be used for classification of the message as sell, buy or hold (more details on the use of General Inquirer's Harvard IV-4 Dictionary will be provided in the following subsection). The following subsection describes the individual sub-components in the first framework component of the design as well as the major role played by each of these sub-components in the overall function of this component.

- **Stock Tweets Data**

The primary data for this study were obtained from Stocktwits.com (<http://www.stocktwits.com>). One year of StockTwits data were downloaded from the website's Application Programming Interface (API) for the period of April 3rd 2012 to April 5th, 2013⁹. The sample period consists of 252 days only because the U.S. stock market is idle at weekends and on national holidays. Over 3,541,959 stock micro-blog posts were obtained from API. StockTwits messages related to the companies making up the Dow Jones Industrial Averages Index were filtered out and returned for this research study along with the required information related to each message, such as user ID, content of the message and the published date and time. A complete list of the required attributes of StockTwits needed for this study can be found in Table 4.4. The StockTwits API Schema, which describes the full StockTwits data, will be provided in Appendix III.

⁹ In order to download the StockTwits data, a StockTwits agreement form was signed by the Head of Contract and IP of the research support and development office of Brunel University and the StockTwits website. The signed Licence Agreement is provided in Appendix II.

Table 4.4: The list of the required attributes for StockTwits collection

StockTwits Data	Attributes for collection
ID	StockTwits unique identifier for the message
Body	Message content
Created_at	Date and time when the message was created

Table 4.5 shows a few typical examples of the StockTwits messages, which are presented in their original format before pre-processing.

Table 4.5: Examples of StockTwits messages

ID	Tweet	Date	Time
12488749	"\$IBM out half +.50"	11/03/2013	17:30:13
9901572	"\$INTC short from Thursday working well. Up 2% with it so far. http://stks.co/mC9s "	09/10/2012	17:12:31
9611602	"\$MA \$V \$AXP just wait until mobile payments overtake cash"	20/09/2012	15:46:30
12158099	"\$VZ breaking out through 45 level with volume"	20/02/2013	18:20:52
7503061	"The Cramer on \$INTC and \$MSFT: http://stks.co/3EI2 (holding both)"	05/04/2012	01:04:16
11147935	"\$JPM - Buy 43.50 puts for next week."	22/12/2012	13:11:07
12514291	"Current holdings: \$ADP \$T \$V \$ERX \$XLU \$QCOM \$MSFT \$ALTR \$MUR"	13/03/2013	20:39:09
9562805	"\$GS looks good here.....&122 YOU PRESS LONGgot bull flag? http://stks.co/fBHM "	17/09/2012	22:43:01
10837630	"\$MSFT for long term short!!!!!!!!!"	05/12/2012	07:37:10
10171127	"\$BA Buying before call with good numbers."	24/10/2012	14:06:15
11677420	"\$DIS bearish to downside to 51.50"	25/01/2013	02:57:41
9541300	"\$UNH vs. KFT News ~ Dow Swaps Out Kraft for United health ~ http://stks.co/iB6k "	15/09/2012	23:35:36

- **General Inquirer’s Harvard IV-4 Dictionary**

General Inquirer is a well-known and widely used program for text analysis. The Harvard IV-4 Classification Dictionary on the General Inquirer’s website lists each word as either positive or negative¹⁰. Many psychological finance studies have

¹⁰ The Harvard IV-4 Dictionary contains more than 4000 emotional words are classified as either positive or negative and are obtained from (<http://www.wjh.harvard.edu/~inquirer/homecat.htm>).

Chapter Four: Research Methodology

used the Harvard IV-4 Dictionary for various text analysis tasks (Tetlock, 2007; Engleberg, 2009; Kothari et al., 2009; Loughran and McDonald, 2011). The General Inquirer's Harvard-IV-4 classification dictionary of emotional words is used in this thesis to add each occurrence of emotional words in a message to the bag of words (Tetlock et al., 2008). From the domain knowledge of Harvard-IV dictionary, more than 4,000 emotional words are tagged and classified as either positive or negative. This builds on the results of Tetlock et al. (2008), who found that fractions of emotional words (negative words) in firm-specific news stories can predict individual firms' accounting earnings and stock returns. Therefore, at this point, text-mining approaches based on a pre-defined dictionary are combined with statistical methods.

A glance at the most commonly occurring words in StockTwits posts provides a reasonable idea of the linguistic bullishness of the three classes (buy, hold, or sell). Since a bull message indicates that an investor is optimistic and sends a "buy" signal to the market participants, it is therefore likely to associate positive emotions with the "buy" class. On the other hand, when an investor posts a bear message, this indicates that the investor is pessimistic, sending a "sell" signal to other market participants. Since sell signals contain many bearish words, it is therefore important to associate negative emotions with the "sell" class. This supports the findings of Tetlock et al. (2008) that negative words are among the most common features of "sell" signals. The "hold" class is more likely to contain an equal balance of positive and negative emotions. It also contains neutral words such as the name of the company or product names.

- **Financial Market Data**

The financial data are obtained from Bloomberg for the actively traded blue chip stocks of the 30 companies making up the DJIA index as well as the data for the DJIA index itself, for the period between April 3rd 2012 and April 5th 2013. No extraordinary market conditions were reported during this period, so it represents a good base test for the evaluation. At daily intervals, the price data on high, low, opening and closing stages of the day; the trading volumes for all 30 stocks and the index are obtained over the same period of time. These daily prices and trading

Chapter Four: Research Methodology

volumes will then be used to calculate other financial variables in this research study (e.g. return and volatility).

- **Text Preparation**

Text preparation is considered the initial stage of the textual data mining process. At this stage, pre-processing of textual data is carried out and the selection of input variables or attributes should be identified. The task of selecting input variables (a so-called “bag of words” approach using the feature selection method) needs to be interactively and collaboratively determined by data mining and human experts (e.g. financial managers) in the domain field of data (financial data). The guidance of domain experts can help determine which terms or phrases are more appropriate in textual analysis. These input variables must then be coded and put in a format suitable for text data-mining (TDM) tasks. In this research study, the feature selection approach will be used as the data-mining tool to select input variables and to extract relevant features from the datasets. The following section will elaborate on the feature selection process and the different methods used to perform the feature selection task.

The next step at this stage of the analysis is to apply some pre-processing techniques. **Six** pre-processing steps are performed to improve the quality of data input and reduce feature space. The first step is to remove the unnecessary words or noise words with low effectiveness in textual analysis of the data. These words include some verbs (e.g. is, are, were etc.), pronouns and other words (e.g., “a”, “an”, “the”, or “and” etc.), which are called stop words, need to be removed. The advantage of removing such words is that text is cleansed of the ineffective words and can be interpreted in a more effective and efficient manner. The omission of these less informative words improves the accuracy of results of the text-mining process and is considered a common task in most text-mining applications (Blair, 1979). While unnecessary words are removed from the list, the addition of other words that were relevant to a particular context (e.g. in this research study, company names proved to be relevant) is also effective in textual data analysis. Second, text reformatting needs to be performed (e.g. whitespace removal). Third, all tweet data should be converted to lower-case characters. The assumption behind this is that an automated algorithm might treat any of these characters separately (e.g. “sell” and “Sell” would be two distinct features). Fourth, the most widely used Porter stemmer approach is applied for the purpose of removing suffixes or (morphological endings) from words. Word

Chapter Four: Research Methodology

stemming is one of the important pre-processing steps to consider. It refers to the process of bringing words back to their actual form. In other words, it is the process of shortening derived words to their initial roots. For example, words such as “buys” and “buying” are stemmed to their base word “buy” (Porter, 1980). Fifth, tokenisation must be performed on the database; this can be defined as a process of replacing all values, symbols, percentages, hyperlinks and figures with a token (text). For example, all stock tickers “\$ticker” of the companies are replaced by the token (“Stocksign”), the characters “\$\$” or “\$\$\$”, which are most commonly used as abbreviations for the term “money”, are replaced by a common format (“money”) and the @ sign in the tweets is replaced by text (“at”). Sixth, all duplicated tweets by the same user and those tweets posted over the weekends and on public holidays were removed.

- **StockTwits Sentiments Manual Labelling**

A random selection from a representative sample of 2,892 of tweets from all 30 stocks on the Dow Jones Index is hand-labelled as a “buy”, “hold” or “sell” signal. These hand-labelled messages constitute the training set which is then used as an input for the model of different machine learning algorithms. As discussed in the previous section, for manual classification the researcher depends heavily on the Harvard-IV dictionary by looking at the most common words frequently appearing in postings that provide reflections of the linguistic bullishness of each of the three distinct classes (buy, hold or sell). The general rules that are applied when labelling the data are provided in the coding Scheme in Section 4.7.2.

4.8.2 Feature Selection and Construction Framework

Feature selection is one of the data-mining tools most commonly used to select sets of relevant features from datasets based on some predetermined criteria (Sima and Dougherty, 2008). It is an essential pre-processing step in the data-mining process (Zhang et al., 2008). It relies upon a single assumption about the datasets, i.e. that the subset of features contains relevant and/or irrelevant features. Feature selection aims to limit the effect of irrelevant features by focusing only on useful or relevant features from the original subset (de Souza et al., 2006). Since feature selection concentrates on selecting the relevant features and omitting irrelevancies, it makes data-mining tasks easier while enhancing the ability to reveal relevancies within the data (Czekaj

Chapter Four: Research Methodology

et al., 2008). It also results in a high prediction accuracy in the classification problem (Guyon and Elisseeff, 2003; Yang and Olafsson, 2006).

The concept of feature relevancy was formalised by Kohavi and John (1997) and implies that features in a given text or a documents should be classified into three categories: strongly relevant, weakly relevant, and irrelevant. The strongly relevant features are those that hold useful information, that do not exist in any other combination of features and whose removal would cause a definite loss of prediction accuracy of a given classifier. The weakly relevant features are those that contain information that either occurs in the combination of strongly relevant features or is already present in other weakly relevant features. The weakly relevant features sometimes contribute to the prediction accuracy. However, the irrelevant features are removed because they contain no useful information about the classification problem (Blum and Langley, 1997). The irrelevant features are those that do not contribute to the prediction accuracy.

Feature selection can formally be defined by introducing the following notation. Suppose F is the given subset of original features with n numbers of features in subset F . Let \bar{F} denote the selected features with \bar{n} numbers of features in subset \bar{F} where $\bar{F} \subseteq F$. In this case, the criterion rule for selecting the subset \bar{F} from subset feature F will be denoted by $J(\bar{F})$. Therefore, in accordance with the basic assumption of feature selection in selecting the most relevant features, the higher the value of $J(\bar{F})$ the more relevant the feature. Consequently, the aim is to maximise the value of $J(Z)$ where Z is the most relevant subset feature in feature set \bar{F} . The feature selection will typically be defined in the following expression:

$$J(\bar{F}) \max_{Z \subseteq \bar{F}, |Z|=n} J(Z) \quad (4.1)$$

The feature subset that maximises $J(Z)$ will be achieved through the feature selection process. Feature selection is a process that typically involves four basic steps (Liu and Yu, 2005) as shown in Figure 4.2. These steps are subset generation, subset evaluation, stopping criterion and results validation. The subset generation is the first step in the feature selection process in which the subset features are produced based on a certain search strategy. In the second step (subset evaluation), each subset value

Chapter Four: Research Methodology

is then evaluated by comparing it with the previous optimal value of subset feature according to a predefined evaluation criterion. If a feature reports a new best (optimal) value, then it replaces the previous optimal and it will be the basis of comparison of the following subset feature generated from the first step. The first and second steps (subset generation and subset evaluation) therefore continue until a given stopping criterion is reached. Either the subset generation or subset evaluation will determine the stopping criterion. The feature selection process ends with the validation of results, where the selected best subset features are subject to a validation test. The validity of the selected features is determined by carrying out different tests and comparing the results with the previously established results.

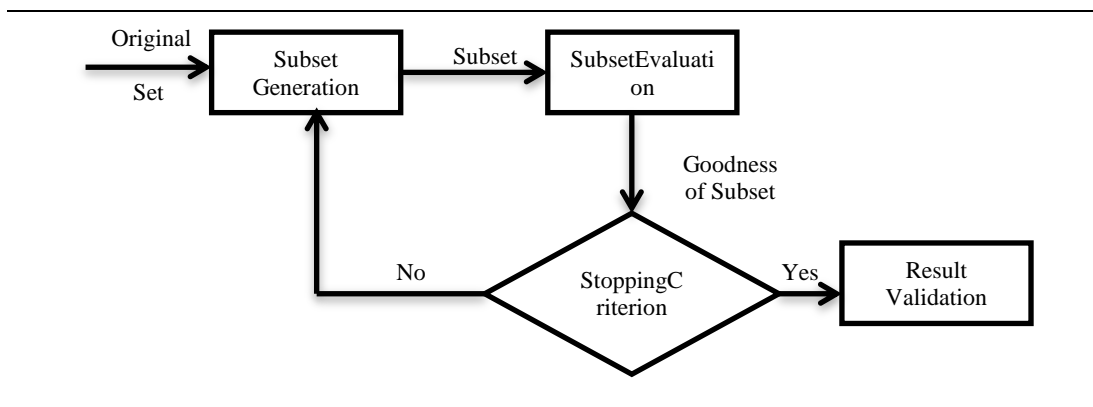


Figure 4.2: The feature selection process
Source; Adopted from (Liu and Yu, 2005)

- **Filter Approach**

The filter method is typically the initial approach to feature selection. It evaluates the relevance of features by using the intrinsic properties of the training set. It employs some statistical measures (Li et al., 2009) to decide about the appropriate features and which to retain or remove. Filter selection procedures can be described in four steps as shown in Figure 4.3. First, as with any feature selection process, it is necessary to decide on the search strategy and determine the direction of the search in order to generate and produce the relevant features in the datasets. Second, for every selected feature produced from the first step, a relevant score based on a statistical measure (Liu and Yu, 2005) will be assigned (either high or low). Third, the assigned features will then be arranged in a list in accordance with their relevancy values, where the features with high relevancy values will be at the top and the features with low values will be at the bottom of the list. Sometimes, however, the classifier will

Chapter Four: Research Methodology

return only the highly relevant features that are considered informative while the non-informative features with low values will be discarded (Sayes et al., 2007). Fourth, the selected features (high/best relevance features) will then be obtained and fed as inputs into the machine learning system (classifier). Finally, these selected features are evaluated by the accuracy of classifiers using various performance evaluation techniques (more details about the performance evaluation will be introduced in Section 4.6)

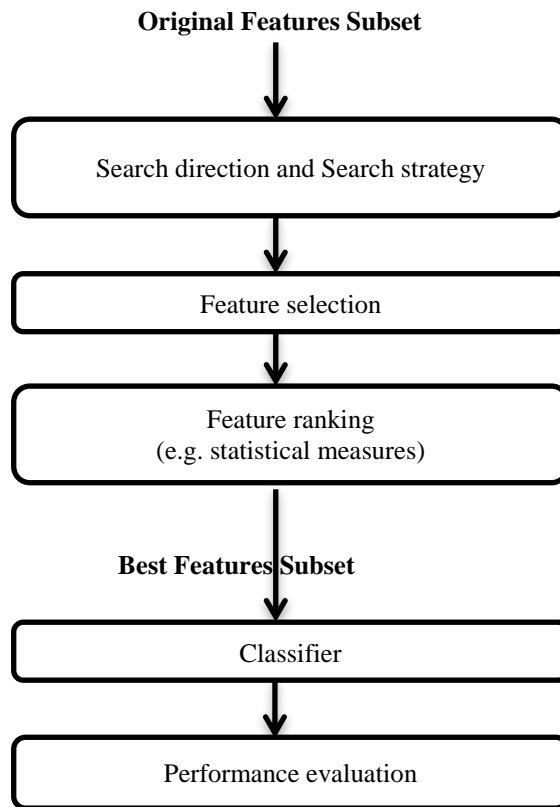


Figure 4.3: The Process of Filter Feature Selection

Information Gain Criteria: Information Gain (IG) is the most commonly employed criterion for evaluating the goodness of the features in a machine-learning environment. It uses *Ranker* as a search method which ranks the attributes by their individual evaluations. Information gain is biased in favour of features with higher dispersion (Huang et al., 2008). IG measures the amount of information obtained for the predicted class within the dataset by perceiving the absence and the presence of a feature (Yu and Liu, 2004). It is calculated based on the following formula:

$$IG(f_k) = \sum_{c \in (c_i, \bar{c}_i)} \sum_{f \in (f_k, \bar{f}_k)} \Pr(f, c) \log \frac{\Pr(f, c)}{\Pr(f) \times \Pr(c)} \quad (4.2)$$

where, f_k means the presence of the features k and $\overline{f_k}$ indicates the absence of feature k . After the attribute selection is performed, a list of all subset attributes along with their relevance rank is shown in the output result. The output results rank attributes based on the relevant statistical score in which the attributes are arranged in accordance with the relevancy value. The top features in the list indicate the high-relevance features, while the low-relevance features are located at the bottom of the list. Performing feature selection by omitting the low-relevance features down the list and retaining the most (best) relevant features will improve the classification accuracy of different machine-learning classifiers.

- **Wrapper Approach**

The Wrapper method evaluates the relevancy of the subset features by choosing the optimal relevant features from the original datasets through the use of a special classifier as the evaluation criterion. This means that the optimal features selected under the Wrapper approach are tailored to a particular classifier and may not be applicable in any other machine-learning system. This may be due to the fact that each classifier has a different bias that might have different effects on the selection process. The term “Wrapper” comes from the fact that the feature selection process is “wrapped” around a particular classifier. Therefore, the classifier plays an important role in the Wrapper approach. Figure 4.4 shows the process of Wrapper feature selection. As with the Filter approach, the feature subsets are produced in the generation steps through the use of a search strategy and direction. As exhibited in Figure 4.4 Wrapper methods repeatedly call the classifier to be run on the subset features and must be re-run when different subsets from the original features are produced. This is a time-intensive process because of the repetitive engagement of the classification algorithm for all the subset features in the datasets. To evaluate the accuracy of each subset’s features on the training data, the estimated accuracy of the classifier (e.g. cross-validation) must be used (John et al., 1994). Then, the features with the highest accuracy rates will be chosen as inputs fed into the classifier algorithm. The features with low accuracy rates will be removed and the classifier will then be called to re-run on new subset features from the original feature sets.

Chapter Four: Research Methodology

The inverted rows in Figure 4.4 indicates the repetitive procedures of the Wrapper method for different subset features.

The Wrapper approach can be applied to any type of machine-learning algorithm. There are three types of classifiers most widely used in the Wrapper approach for feature selection purposes: Decision Trees Classifier, Naive Bayes algorithm and Support Vector Machine. These types of classifiers will be addressed in this thesis for the feature selection purpose. Different areas of research including Web mining and financial analysis have received fruitful attention that made use of these classifiers to perform Wrapper feature selection. The focus of this thesis is to use the three classifiers' techniques in feature selection for financial prediction application and analysis.

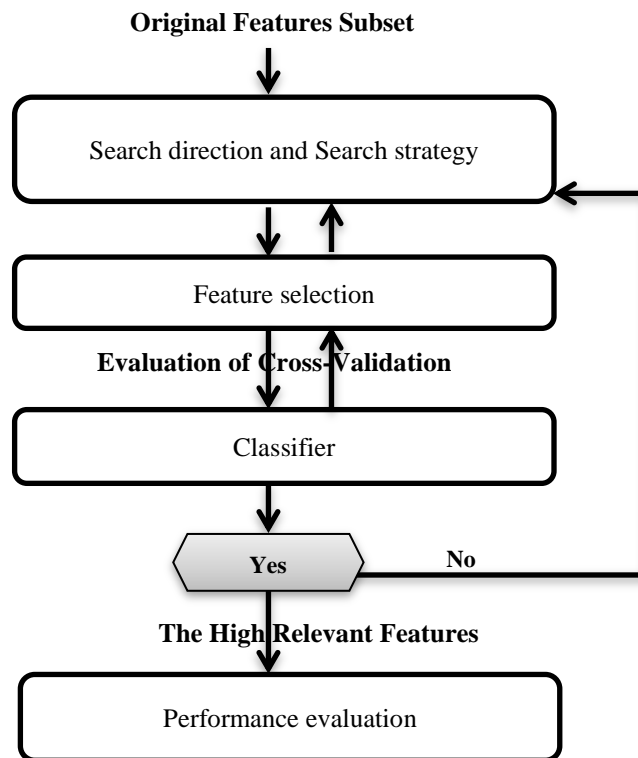


Figure 4.4: The Process of Wrapper Feature Selection

In summary, Filter and Wrapper methods have both been used in many research studies (e.g., Inza et al, 2004; Ruiz et al, 2006; Zheng and Zhang, 2008) for the purpose of feature selection. It has been found that Wrapper methods often perform better in terms of classification accuracy than the Filter approach. However,

Chapter Four: Research Methodology

the optimal features selected in the Wrapper method tend to be tailored to a particular classifier. The main issue is that each classifier has its own biases and nature that may result in different optimal features being selected under each of these classifiers.

4.8.3 Text-Processing Models

After the initial pre-processing stages have been completed, the next essential stage of the text-mining process is text processing. In the text-processing stage, the information will be structured, organised and stored in a formatted structure for further analysis. At this stage of analysis, an appropriate data-mining technique is selected which is used to process the data and help optimise the results. Meaningful features are then extracted through the application of some structural techniques available in the literature such as Decision Trees, Naive Bayes classifier and Support Vector Machine. The following section will elaborate in more detail the three models of machine-learning classifiers, i.e. Naive Bayes, Decision Tree and Support Vector Machine, which are applied to the sentiment detection process in this thesis.

- **Naive Bayes Classifier**

A Naive Bayes classifier is a simple classifier technique based on the Naive Bayes Theorem. It is a well-known approach most commonly used in solving practical domain problems. It is based on the assumption called the Naive assumption which states that a given attribute is independent of the other attributes contained in a given sample, and it considers each of these attributes discretely when classifying a new incoming instance. The Naive Bayes algorithm is based on the joined probabilities of words or a document belonging to a class in a given text (Witten et al., 1999). The probability of a document d belonging to class c is calculated based on the Bayes rule by the following formula:

$$P(c | d) = \ln P(c) \sum_{1 \leq i \leq n_d} \ln P(w_i | c) \quad (4.3)$$

where $P(c)$ is the prior probability of a document belonging to a class c . $P(w_i | c)$ is the class-conditional probability of word w_i occurring in a document of class c . Ln is the natural algorithm used to assign the document to the class which represents the “naive”

Chapter Four: Research Methodology

assumption that the occurrence of words or attributes are independent of each other. Both probabilities $p(c)$ and $P(w_i|c)$ are estimated based on manually coded documents (tweets) of the training set. Therefore, the prior probability is computed as follows:

$$\hat{P}(c) = \frac{N_c}{N} \quad (4.4)$$

where N_c refers to the number of documents or document frequency in class c , and N is the total number of documents. The class conditional probability $P(w_i | c)$ is estimated and calculated based on the following formula:

$$\hat{P}(w_i | c) = \frac{W_c}{\sum W_c} \quad (4.5)$$

Where W_c is the total number of words w in a given document of class c .

- **Decision Tree Classifier**

The decision tree method is one of the most frequently used techniques for classification problems. It is a tree structure consisting of nodes, leaves and branches. Decision trees used for classification problems are often called classification trees where each node represents the predicted class of a given feature. They are also used for regression problems where each node is indicated by an equation to identify the predicted value of an input feature. It applies the concept of information gain or entropy reduction, which is based on the selection of a decision node and further splitting the nodes into sub-nodes. This function is performed by building decision trees (or decision nodes) from a set of training data. In this research study the training data are a set of sample classes to which each classified tweet belongs (e.g. $C = c_1, c_2, c_3, \dots, c_n$). The tweets that have already classified $T = t_1, t_2, t_3 \dots t_n$ consist of different attributes or features ' x ' of the so-called vector (e.g. $t_1 = x_1, x_2, x_3, \dots, x_n$).

A decision tree algorithm C 4.5 is an extension of Quinlan's algorithm ID3 that generates decision trees or nodes (Quinlan, 1993) by choosing the most effective attribute that splits each node into sub-nodes augmented in one class or another. The normalised information gain is an impurity-based criterion that uses the entropy measure (Rokach and Maimon, 2005) to evaluate the effectiveness of an attribute for

Chapter Four: Research Methodology

splitting the data. Therefore, these criteria state that the attribute with the greatest normalised information gain is chosen to make the decision. The process of splitting the decision nodes continues until no further split is possible. This means that the data have been classified as close to perfection as possible. This process safeguards maximum accuracy in the training data. To form a decision tree, the following steps are required:

Step 1: Define the entropy of x

$$H(x) = - \sum_i^k P_i \log_2(P_i), \quad (4.6)$$

where x is a random variable with k discrete values, distributed according to probability value $P = (P_1, P_2, P_3, \dots, P_n)$ of class subset i .

Step 2: Calculate the weighted sum of the entropies for each subset.

$$H_s(T) = \sum_{i=1}^k P_i H_s(T_i), \quad (4.7)$$

Where P_i is the proportion of records in subset i .

Step 3: Calculate the information gain

$$\text{Information gain } IG(S) = H(T) - H_s(T) \quad (4.8)$$

The information gain is the criteria necessary to choose the most effective attribute to make the decision. Then the selection of attribute at each decision node will be the one with the highest information gain, $IG(S)$.

- **Support Vector Machines (SVMs)**

Support vector machines (SVM) are the most widely used techniques for textual analysis applications; they have proven excellent empirical success with strong theoretical foundations (Tong and Koller, 2002). They were first developed in Russia in the 1960s (Vapnik and Lerner 1963; Vapnik and Chervonenkis 1964). Compared with traditional methods, which minimise empirical training errors, SVMs implement the structural risk minimisation principle (SRM) (Cho et al., 2005; Lin et al., 2006). SVMs aim to minimise the upper bound of the generalised error via the optimal margin between separating the hyperplane and the data (Amari and Wu,

Chapter Four: Research Methodology

1999).

The primary aim of SVM is to find a maximum hyperplane, which clearly separates the instances and non-instances of a given class relative to the target variables (Barakat and Bradley, 2007). This common approach is generally used when the instances of the target variables are described as linearly separable whereby the target variable should have only two class values. On the other hand, there are some cases where the target variable may have more than two class values where the instances assigned to these class values are, in this case, described as non linearly separable. With non-linearly separable data it is hard to find an optimal hyperplane to classify the data instances. Therefore, for a non-linearly separable data, SVM makes use of Kernel methods to transform the data from an input space or parametric space into a high-dimensional feature space. According to the Mercer theorem (Vapnik, 1998), the Kernel function implicitly maps the data, that are linearly non-separable, into a linear separable from input vector to high-dimensional feature space (Figure 4.5). Therefore, the non-linear separable data in parametric space ϕ could be extended to a linear separable (Aizerman et al., 1964) by adequately mapping the data from the input space $S = \{x\}$ into possibly a high dimensional feature space $F = \{\phi(x)\}$ (Amari and Wu, 1999; Lu et al., 2003). Different Kernel functions are used to map the non-linear separable data points into high-dimensional space to map the data to linear separable. However, the domain of the study and the type of data examined specify the choice of Kernel. A linear Kernel function would be the best choice in the context of this research study based on statistical textual analysis.

SVMs were generally used for two classification problems: binary classification and multi-class classification. SVMs are also used in regression and time-series prediction applications, and they have delivered excellent performances (Smola and Scholkope, 2004). Since the primary concern of this study is to predict the real value of stock prices, the use of Support Vector Regression algorithm (SVR) might be well suited to textual analysis of StockTwits. SVR has also been well documented in time-series forecasting applications such as in the works by Mukherjee et al. (1997), Thissen et al. (2003) and Muller et al. (1997). Kim (2003) proposed an SVM approach to predict the direction of stock prices.

The SVM model represents the instances in a class as points in space,

Chapter Four: Research Methodology

adequately mapped so that instances of other classes (represented as points) in the same space are widely separated and depicted as widely as possible (Figure 4.5). New instances are then plotted in the same space and the classes to which these instances belong will depend on which side of the space they fall. The Kernel function will therefore arrange the instances within the multi-dimensional space by using a hyperplane to separate the data instances of two classes of patterns (Amari and Wu, 1999).

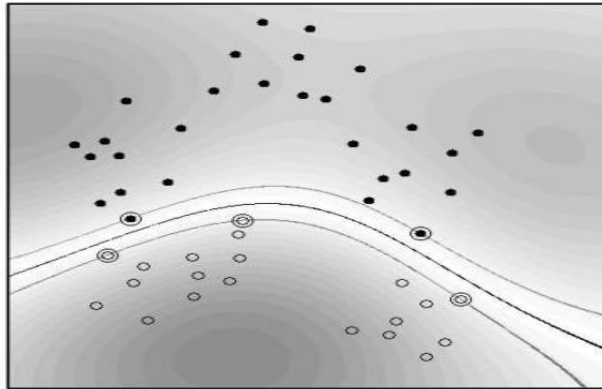


Figure 4.5: The maximum hyper-plan of Support Vector Machine

This Figure shows the maximum hyperplane (the optimal margin) with its support vectors of Kernel k . The support vectors with double circles are an indication of the vectors on the margin line (Chen et al., 2005).

SVM attempts to maximise the margin space (which is denoted by $2/\|\mathbf{w}\|$) between the separating hyperplane and the data instances (Figure 4.5). Such instances are known as vectors of Kernel K that are engaged in building the support vector model for generalisation. The margin, therefore, will be the measure of the generalisation ability of the hyperplanes to separate the data instances into the corresponding classes. The larger the margin, the better the generalisation abilities are expected to be (Christiniani and Shawe-Taylor 2000).

Consider the two classes of a training dataset given as:

$$D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\} \subset \mathbb{R}^n \times \mathbb{R} \quad (4.9)$$

In the linear regression of SVM, they are estimated in the following function:

$$f(x) = (\mathbf{w}, \mathbf{x}) + b \quad \mathbf{w}, \mathbf{x} \in \mathbb{R}^n, b \in \mathbb{R} \quad (4.10)$$

By minimising the regularised risk function as stated in Cortes and Vapnik (1995).

$$\min \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*) \quad (4.11)$$

Chapter Four: Research Methodology

$$\text{subject to } \begin{cases} y_i - (w, x_i) - b & \leq \varepsilon + \xi_i \\ (w, x_i) + b - y_i & \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* & \geq 0 \end{cases} \quad (4.12)$$

Minimising the regularised term $\frac{1}{2} \|w\|^2$ will make the function as flat as possible and that will play a major role in controlling the function capacity (Lin et al., 2006). The term $(\xi_i + \xi_i^*)$ is the empirical error measured by the loss function. The constant $C > 0$ measures the flatness of the function f and determines the maximum value of tolerated deviation from the loss function ε . It is also called the regularisation constant that determines the trade-off between the flatness of the function and the deviation from ε (Smola and Scholkope, 2004). The ε -insensitive loss function, which is analogous to the “soft margin” (Bennett and Mangasarian 1992), is described by:

$$|\xi|_\varepsilon = \begin{cases} 0 & \text{if } |\xi| \leq \varepsilon \\ |\xi| - \varepsilon & \text{otherwise,} \end{cases} \quad (4.13)$$

From the equation given in (4.9) the estimated weight vectors w is defined as

$$\bar{w} = \sum_{i=1}^n \beta_i x_i \quad (4.14)$$

And,

$$\bar{b} = -\frac{1}{2} \{w_i, (x_r + x_s)\} \quad (4.15)$$

where β are the coefficients of the samples. In SVR the β coefficients of the support vectors should not be equal to zero. Figure 4.6 illustrates the situation graphically, with the dotted tube representing the loss function ε . In such a case, where the predicted value falls in the dotted tube the loss is zero, while if the predicted value is outside the tube, it contributes to the loss. The loss is penalised in a linear fashion by calculating the distance between the predicted point and the radius ε of a given tube.

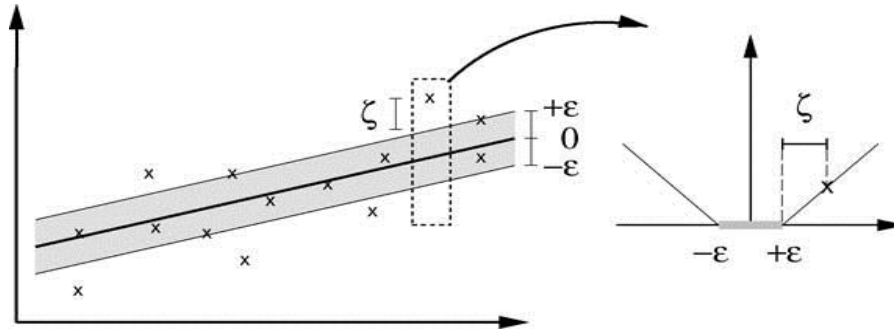


Figure 4.6: The soft margin loss and ϵ -insensitive loss function for a linear SVM
Source: Adopted from (Scholopf and Smola, 2002).

- **Justification for Choosing the Classifier Model**

In the present research, different data-mining approaches are used for the textual classification task. A benchmark of the models (which will be discussed later in this chapter in the performance evaluation framework) is used to test the validity and accuracy of each model to determine which model is the most suitable for sentiment classification of StockTwits data. This study has focused on the application of three models: Decision Tree, Naive Bayes Algorithm and Support Vector Machine (SVM). There are several reasons for the adoption of these different classifier techniques. The main reason is that these classifiers adopt various selection criteria for classifying the data variables. For example, in the Decision Tree, the information variables are selected based on entropy measures. Meanwhile, the Naive Bayes algorithm uses probabilistic information selection criteria, and the SVM makes use of Kernel functions to map and select the input variables. Another reason is the distinct features and advantages of each individual classification method.

Naive Bayes (NB) classifier is accurate, time-efficient and the simplest method of implementation, all of which are attractive reasons for choosing this classifier technique. The Naive assumption of the independent occurrence of each attribute adds value to the Naive Bayes technique where the importance of each of these attributes is considered equally likely. Another distinct feature of the NB classifier is its use of all attributes regardless of the size of the dataset being considered; thus, it overcomes the problem of handling missing values.

Decision Trees classifier is the most widely used approach for textual analysis. It has been successfully applied in any field that requires any form of data mining and textual analysis. The unique ability to handle large databases containing hundreds or thousands of features makes the decision tree classifier superior to other data-mining

Chapter Four: Research Methodology

techniques. Moreover, the non-parametric feature of most decision tree algorithms makes it very easy to understand as it does not require an expert in the field of data that is being mined. The rule created for each path in the tree from the node to a leaf node is in the form of the “If - Then” rule which makes DT simple and easier to understand (Chien et al., 2007)

Support Vector Machines (SVMs) are powerful classification techniques that have been widely used in text classification tasks. In practice, SVM classifiers have proved successful and have been found superior when other classifiers have performed poorly. They are considered attractive classifiers as they can handle both linear and non-linear classification problems when other classifiers fail to do. Due to this fact, SVMs have proved empirically successful in performing feature selection tasks in different kinds of databases (e.g. genes and webpages). They are efficient classifiers in terms of both computational time processing and complexity of datasets involved.

Having justified the reasons for the use of different classifiers employed in this research, it is also important to note some of the drawbacks associated with each of them. Table 4.6 summarises the advantages and disadvantages of each of the three different classifiers (Naive Bayes, Decision Tree and Support Vector Machine).

Table 4.6: A list of advantages and disadvantages of Naive Bayes, Decision Tree and Support Vector Machine Classifiers

Classifier	Advantages	Disadvantages
Naive Bayes (Witten et al., 1999)	<ul style="list-style-type: none">- Simple, accurate, fast and easy to implement for data mining and textual analysis application.- Overcoming the problem of handling missing data in large datasets-The “Naive assumption” adds value to Naive Bayes technique where the occurrence of each attribute is independent of each other.- The efficiency in processing time.	<ul style="list-style-type: none">- The inability to handle the data as one stream; rather, it divides the dataset in different ranges or classes, which may ultimately affect the results.
Decision Tree	<ul style="list-style-type: none">- Simple, easy to interpret and explain.- Handles large databases while the	<ul style="list-style-type: none">- Despite the simplicity of the rule generated by decision tree

Chapter Four: Research Methodology

(Quinlan, 1992)	<p>decision nodes and decision tree's root built from the datasets are independent of its size.</p> <ul style="list-style-type: none"> - The optimal selection split through the process of recursive classification ensures maximum generalisation and high accuracy in the training data - Comprehensibility of discovery knowledge, which, is measured by the number of leaves in the composite classifier. 	<p>methods, the branches built can be very large and extended and become difficult to interpret.</p> <ul style="list-style-type: none"> - The continuous process of classification of the data and on-going process of building the decision trees generate time-complexity problems, especially large datasets.
<p>Support Vector Machine (Cho et al. 2005)</p>	<ul style="list-style-type: none"> - Good Generalisation performance: the rules generated in SVM are easily learnt to correctly classify a new instance in any given training sample. - Computational Efficiency: SVMs are efficient in terms of processing time and complexity of data involved. - Robust in high dimensions: SVMs perform well in high-dimensional data and have the ability to overcome the problem of over-fitting. 	<ul style="list-style-type: none"> - The main limitation of SVM is the model building time, due to the quadratic nature of the algorithm for building an SVM. - The use of different Kernel functions to transform the non-linear separable data points into linear separable causes difficulty in interpreting the model.

Source: Adopted with modification from Ur-Rahman, (2010)

4.8.4 Performance Evaluation

This is a decision-making stage where the effectiveness of the models used in the text-processing model framework is tested in order to select the best model. This section presents the performance evaluation methods used to evaluate the different models of text-mining techniques adopted in this thesis. In machine learning, there are many methods for estimating the quality of classification algorithms. In this thesis four methods of performance evaluation are used to assess the classifier's quality and effectiveness. These methods are as follows: confusion matrix; classification accuracy and error rate; analysis of sensitivity and specificity; and k-fold cross-validation. These methods are explained in the following subsections.

- **Confusion Matrix**

The confusion matrix analysis is regarded as the most direct and significant way of measuring the quality and performance of the classifiers' algorithms. The instance outputs produced by the classifier algorithm during the testing stage are

Chapter Four: Research Methodology

generally tallied and tested for correct and incorrect classification of each class label (Bradley, 1997). These data instances are then displayed in a confusion matrix. A confusion matrix is a form of generic contingency table that contains information about the actual and predicted classes for a set of labeled data (Polat and Güneş, 2007). Data displayed in the confusion matrix contain all of the information needed to evaluate the classifier's performance. Table 4.7 illustrates the confusion matrix for a binary class classifier.

Table 4.7: Representation of confusion matrix		
Actual	Predicted	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

The entries in the confusion matrix are explained as follows:

- TP is the number of correct predictions that an instance is positive,
- FN is the number of incorrect predictions that an instance is negative,
- FP is the number of incorrect predictions that an instance is positive,
- TN is the number of correct predictions that an instance is negative.

The confusion matrix illustrated in Table 4.7 is held as a baseline where several measurements (e.g. classification accuracy and sensitivity analysis (trade-off between sensitivity and specificity)) can be carried out to evaluate the performance of the classifier algorithm. The following sections will discuss these performance evaluation measurements in more detail.

- **Classification Accuracy**

Classification accuracy is one of the most popular measures for evaluating a classifier system's performance and prediction accuracy (Tan and Gilbert, 2003). It measures the proportion of correctly classified instances in the test set. Error rate is also widely used for measuring a classifier's performance. Both classifier accuracy and error rate are calculated by using values from both lines of the confusion matrix (Prati et al., 2004). From the confusion matrix presented in Table 4.7, the following equation is used to calculate the classification accuracy and error of the test set:

Chapter Four: Research Methodology

$$\text{Accuracy Rate} = \frac{TP+TN}{TP+TN+FP+FN} (\%) \quad (4.16)$$

$$\text{Error Rate} = \frac{FP+FN}{TP+TN+FP+FN} (\%) \quad (4.17)$$

In a normal setting, machine learning algorithms are designed for the purpose of maximising the classification accuracy and minimising the error rate (Kukar and Kononenko, 1998).

- **Analysis of Sensitivity and Specificity**

The correct balance between sensitivity and specificity plays an important role in evaluating classifier performance. The performance of a classifier is evaluated by its accuracy. The accuracy is quantified in the test phase through the calculation of the total number of misclassifications in the test set (Veropoulos et al., 1999). There are two types of misclassifications: false positive and false negative. Evaluating a system's performance is best described in terms of its sensitivity (measuring the fractions of actual positive examples that are correctly classified) and specificity (measuring the fractions of actual negative examples that are correctly classified). A receiver operating characteristics (ROC) graph is a technique used to analyse, compare, organise and select classifiers based on their performance (Fawcett, 2006; Prati et al., 2004). ROC is used to make a comparison analysis based on evaluating the sensitivity and specificity of different classifiers. An ROC graph shows a trade-off between the sensitivity (hit rates) and specificity (false alarm rates) of classifiers (Swets et al., 2000).

For sensitivity and specificity analysis in machine learning, the following definitions and expressions are provided:

The **sensitivity** of a learning machine is defined as the ratio between the numbers of true positive predictions (TP) to the total number of positive instances in the test set:

$$\text{Sensitivity} = \frac{TP}{TP+FN} (\%) \quad (4.18)$$

The **specificity** is defined as the ratio between the number of true negative predictions (TN) and the total number of negative instances in the test set:

$$\text{Specificity} = \frac{TN}{TN+FP} (\%) \quad (4.19)$$

Chapter Four: Research Methodology

where TP, TN, FP and FN denote true positive, true negative, false positive and false negative respectively.

- True positive rate: $TP_{rate} = \frac{TP}{TP+FN}$ is the percentage of positive instances correctly classified and belonging to the positive class.
- True negative rate: $TN_{rate} = \frac{TN}{TN+FP}$ is the percentage of negative instances correctly classified and belonging to the negative class.
- False positive rate: $FP_{rate} = \frac{FP}{FP+TN}$ is the percentage of negative instances misclassified and belonging to the positive class.
- False negative rate: $FN_{rate} = \frac{FN}{FN+TP}$ is the percentage of positive instances misclassified and belong into the negative class.

Spackman (1989) was the first to adopt ROC graphs in machine learning for performance evaluation and comparison among algorithms. ROC is generally a useful performance-visualising method for a classifier (Fawcett, 2006), where the true positive rate (TP_{rate}) is plotted on the y-axis and the false positive rate is plotted on the x-axis (FP_{rate}) (Veropoulos et al., 1999). The ROC curve exemplifies the performance of the classification model through the trade-off between the classifier sensitivity (TP_{rate}) and false alarm rate (FP_{rate}) where the sensitivity can only be increased with a little loss in specificity, and vice versa (Kukar and Kononenko, 1998). The primary aim of a classifier is to minimise the false positive and the false negative rates or, correspondingly, to maximise the true positive and true negative rates (Prati et al., 2004).

- **K-Fold Cross-Validation**

Cross-validation is one of the most important tools commonly used for evaluating classification methods in data-mining applications. It is widely used to predict the generalisation ability of classifier algorithms (i.e. to be generalised to a new example) (Cawley and Talbot, 2003). K-fold cross-validation is one way of evaluating the robustness of the classifier. In k-fold cross-validation, the dataset is partitioned into k subsets. Then, the cross-validation procedure is repeated k times. Each time, one of the k subsets is used once as the test set and the other sets k-1 are combined to form a training set. This repeated process results in k independent

Chapter Four: Research Methodology

realisations of the error measures. The error measures crossing all k trails are then averaged to produce a single estimation (Witten et al., 2011). The advantage of the cross-validation method is that the process is repeated until each subset has had a chance to be a test set exactly once and a training set $k-1$ times (Goldbaum et al., 2002), through which all observations are used for both training and testing with no exceptions. Another advantage is that it is immaterial how the data are divided as the variance of the resulting estimates decrease as the k subset increases (Polat and Gunes, 2007). Moreover, averaging the k error measures acquired across k trails yields an overall error estimate that will normally be more robust than individual measures (Bergmeir and Benitez, 2012).

However, this method also has its drawbacks. Two major drawbacks of k -fold cross-validation are the time inefficiency and extensive computation tasks that the training algorithm has to perform. Under this method the classifier algorithm has to be rerun from scratch k times as it takes k times as much computation to perform the evaluation. Researchers frequently use ten-fold cross-validation as it has proved to be statistically sufficient from the model evaluation method (Witten et al., 1999). In ten-fold cross-validation, the datasets are equally partitioned into ten different subsets. The cross-validation process is repeated ten times; each time, one of the ten subsets will be used as a test set and the other nine subsets will be combined to form the training sets of the model. All ten subsets will have an equal opportunity to be a test set exactly once and a training set nine times. Then, the average error estimates will be calculated across all ten trails.

4.8.5 Training and Testing

The training and testing task is very important in data-mining and machine-learning situations. One of the main objectives of this research study is to automatically detect sentiments in StockTwits messages using various machine-learning techniques. These machine-learning applications are based on different algorithms that are qualified from learning patterns, which fit the primary requirement of this research study precisely.

This study also needs to be able to predict sentiments of further contents of new stock micro-blogging posts (StockTwits messages). In machine-learning situations, once the algorithms are trained on a sample set of data, they are then

Chapter Four: Research Methodology

qualified and capable of predicting any new instance coming from the environment. Therefore, in order to classify/predict any new instances, a model is first built and trained on training data. In most general cases, training data are fed into machine-learning algorithms to produce a classifier which is then tested by an independent test set to produce an evaluation result. The main purpose of the training and testing task is to evaluate whether the classifier can be deployed in a real situation and predict new data coming from the environment. It is very important to note that the training corpus should be kept deliberately small to avoid the problem of over-fitting, which is a common weakness of text-mining algorithms. There are two basic assumptions behind the scenario of training and testing:

- 1- Training and testing sets should be produced from an independent sampling of an infinite population.
- 2- To ensure reliability of the evaluation results, it is very important to ensure that the test set is different from the training set.

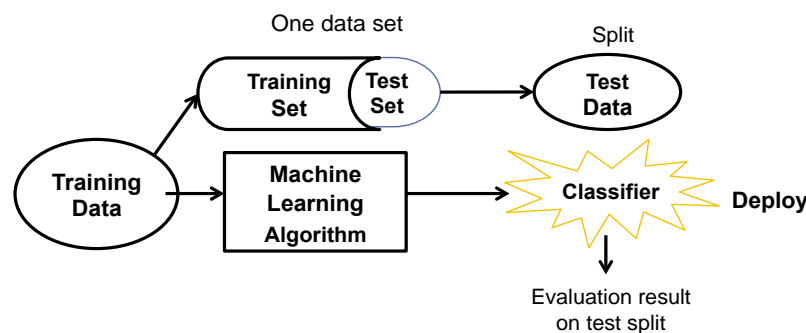


Figure 4.7: Training and testing procedure to assess the model accuracy
Source: Adopted from (Witten, 2013)

Figure 4.7 shows a visualisation of how the training and testing task is generally performed in Weka. There are two methods that are commonly used to perform the training and testing depending on whether one dataset or two separate sets of data are used for training and testing accordingly. Different methods could be used to split the data into training and testing sets. For example, one might train the data on the first ten months of the StockTwits corpus while testing on the remaining two months' corpus. The latter method produces more reliable results when deploying the classifier result to make an accurate prediction of the entire sample of StockTwits data of DJIA in this research. Therefore, in this thesis both methods will be used to

Chapter Four: Research Methodology

ensure the consistency of the results of automatic classification regardless of the methods used to split the data into training and testing sets.

Method (1): Training and Testing Using Automatic Percentage Split (66% Training and 33% Testing) Using One Data Set

In Weka, when one dataset is used, the test will be run automatically by using the percentage split which is set by default; 66% or 2/3 of the total data will be used as the training set while the remaining 34% or 1/3 of the total data will be used as the test set. This method is commonly used when training data are supplied as one dataset. Although this method randomly splits the dataset into training and testing, the results may be misleading. Therefore, to ensure the reliability of the results, training and testing are experimentally repeated at different random seeds (initialising the random number generator to a different amount each time). It is very important to note three basic assumptions used in the repeated training and testing using one dataset:

- 1- Training and testing sets differ from each other and both are independent of an infinite population.
- 2- When using different numbers of random seeds, one should expect slight variations in the accuracy of results.
- 3- Mean and Standard deviations are experimentally calculated from the repeated experimental results using different numbers of seeds. Given these experimental results, the mean and standard deviations are calculated as follows:

$$\text{Sample mean } \bar{x} = \frac{\sum x_i}{n} \quad (4.20)$$

$$\text{Variance } \sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \quad (4.21)$$

$$\text{Standard Deviation } \sigma = \sqrt{\sigma^2} \quad (4.22)$$

Method (2) Training and Testing Using Supplied Test Set of Two Separate Datasets (Training set (In-Sample Set) and Testing Set (Hold-Out Set))

In this method the manually-labelled dataset is divided into two parts: the in-sample set and the hold-out set. In Weka, using the supplied test options, training on the first ten months of the year and testing on the remaining two months generates

Chapter Four: Research Methodology

two separate datasets: training set (in-sample set) and testing set (hold-out set), which are 1,953 and 939 instances respectively. These resulting data are then used for the purpose of the aggregated daily ticker for further analysis. The approach of evaluating the hold-out set is a critical aspect of model fitting where the period of fit (in-sample) is separated from the period of evaluation. This is one of the most reliable methods precisely when the hold-out set is composed of data from a future period where it is used to compare the forecasting accuracy of models' fit to past data (Oh and Sheng, 2011).

4.8.6 Statistical Summary

One of the research questions explored in this study concerns how well stock micro-blogging sentiments can predict stock market behaviour. StockTwits are considered one of the micro-economic indicators that are expected to have an effect on stock market prices. Thus, to answer the above research question, it is necessary to study the relationship between stock micro-blogging features and stock market variables. However, before investigating these relationships, it is important at this stage of the analysis to provide measures for each individual feature of both StockTwits (message volume, bullishness and level of agreement) and Financial Market indicators (trading volume, return and volatility). Figure 4.8 depicts the prospective relationships between tweet features and market features.

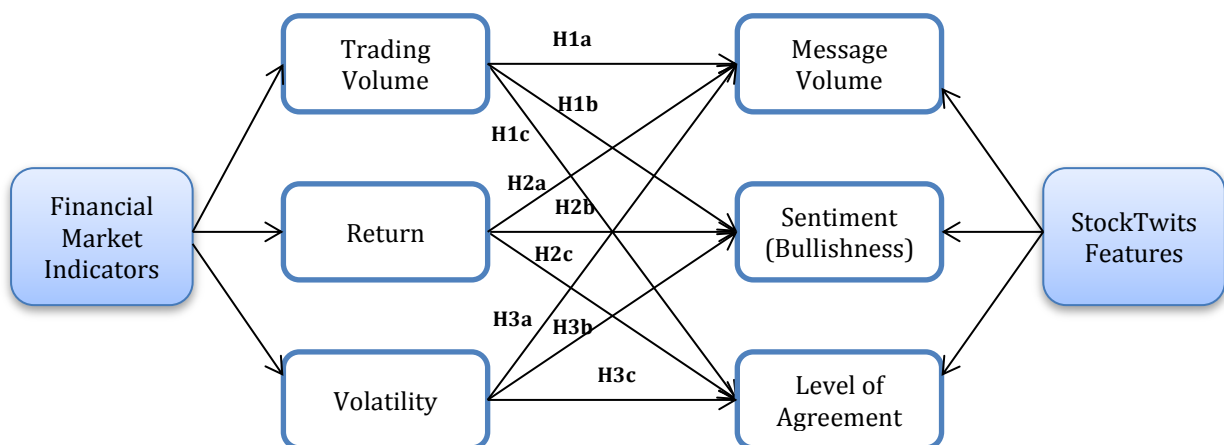


Figure 4.8: The relationship between StockTwits features and Stock Market indicators

Chapter Four: Research Methodology

The following subsections describe how the aggregate of daily tweet features and financial market data is statistically calculated.

A) StockTwits Features

As hundreds of StockTwits messages are posted every day, and in order to study the relationship between tweet messages and market behaviour on a daily basis, tweet features need to be aggregated. This research study focuses on three tweet features: bullishness, message volume and agreement level.

- **Bullishness**

In the stock market, bullishness can be defined as optimism that a particular investment is potentially profitable. For example, when an “investor is bullish (bearish) on stock A”, it means that he/she holds a positive (negative) opinion about the future performance of that particular stock (stock A), thus providing a signal to other investors in the market to buy (sell) more of stock A. Bull investors believe that the market is rising and base their investment strategies on buying more stocks for their portfolios. In contrast, bear investors base their investment strategies on the belief that the market is falling and try to make money by short-selling their investment portfolios.

Sometimes, there appear to be days without any tweets, in which case the silent period is replaced with zeros, following Antweiler and Frank (2004b).¹¹The classification algorithm classified all the tweet messages into three distinct classes M^c where $c \in \{Buy, Hold, Sell\}$. The bullishness of messages is an important tweet feature that determines the proportion of buy and sell signals on a particular day t . This is a measure that is used to aggregate the three different message classes M_t^{Buy} and M_t^{Sell} and M_t^{Hold} in a given time interval¹². The output resulting from the text-processing model framework in Figure 5.1 classifies all the tweet messages into three distinct classes M^c where, $c \in \{Buy, Hold, Sell\}$. The bullishness of messages is an important tweet feature that determines the proportion

¹¹Empirical studies suggested two possible ways to deal with the missing observations in the dataset, either by replacing the missing period with the medians of the respective measures or by filling those missing values with zeros.

¹²All three bullishness measures exclude the number of messages expressing the hold sentiment M_t^{Hold} . The reason for excluding the hold messages is that this type of message holds neutral opinions and thus has no effect on the bullishness measures. Moreover, in most cases this set of messages may contain some amount of “noise” that may bias and distort bullishness signals (Antweiler and Frank, 2004b).

Chapter Four: Research Methodology

of buy and sell signals on a particular day t . “Bullishness” is the measure that is used to aggregate the three different message classes M_t^{Buy} and M_t^{Sell} and M_t^{Hold} in a given time interval. This research study has carried forward the work of Antweiler and Frank, (2004b) by defining bullishness (B_t) using three different measures as follows:

$$B_t = \left[\frac{M_t^{Buy} - M_t^{Sell}}{M_t^{Buy} + M_t^{Sell}} \right] \quad (4.23)$$

$$B_t^* = \ln \left[\frac{1 + M_t^{Buy}}{1 + M_t^{Sell}} \right] = B_t \ln(1 + M_t) \quad (4.24)$$

$$B_t^{**} = (M_t^{Buy} - M_t^{Sell}) = B_t M_t \quad (4.25)$$

where M_t^{Buy} and M_t^{Sell} indicate the total number of traders’ messages conveying buy and sell signals on day t respectively. The first bullishness measure is an essential component for obtaining results of the two other measures while these last two measures are more comprehensive measures as both take into account the number of messages M_t as well as the ratio of bullish to bearish messages. The measure B_t^{**} appears to outperform both alternatives; hence, this measure is used to measure bullishness, which is used as a proxy for investor sentiment in this research study¹³. Because a markedly large number of messages are tweeted on a daily basis, normalisation is therefore needed for these messages as this will assist the model’s estimation. More specifically, as B_t^{**} may contain negative values and in order to take into account such values, the following formula of normalisation is considered:

$$\overline{B}_{it}^{**} = \frac{(B_{it}^{**} - \min B_i^{**})}{(\min B_i^{**} - \max B_i^{**})} \quad (4.26)$$

where \overline{B}_{it}^{**} is the normalised value of bullishness B^{**} of company i at time t , and $\max B_i^{**}$ and $\min B_i^{**}$ indicate respectively the maximum and minimum value of the bullishness measures of company i over the sample period.¹⁴ Note that the normalised

¹³ All analyses of this study is conducted with all three measures of bullishness, and the findings reveals that the third measure $B_t^{**} = (M_t^{Buy} - M_t^{Sell}) = B_t M_t$ outperforms the other two measures; thus only report these results. The reason why B_t^{**} is more robust is the fact that our data were more balanced in terms of the distributions of buy vs. sell messages than Internet message boards.

¹⁴The $\max B_i^{**}$ and $\min B_i^{**}$ of bullishness measures will be different for each company of the DJIA index in the panel series.

Chapter Four: Research Methodology

bullishness is homogenous of a degree between zero and one, in line with the bullishness measure used by Antweiler and Frank (2004b). In addition, our bullishness measure is similar to the investor sentiment index of Wang (2001), who proxies investor sentiment by different types of traders taking into account the minimum and maximum aggregated positions of traders' sentiment.

Since the 'buy' and 'sell' messages indicate that an investor is being bullish and bearish respectively, it is likely that the 'buy' message will be associated with a bullish investor whereas the 'sell' message will be associated with a bearish one. The bullishness index is then computed at the end of each day as a ratio of the number of bullish messages relative to the total number of messages that are either bullish or bearish. This measure represents the number of investors' messages expressing a particular sentiment (buy or sell), giving more weight to a larger number of messages in a specific sentiment. Because the conversations taking place in the StockTwits forums target the individual investors in the stock market, the sentiment made by these platforms influences the trading decisions of such investors and hence serves as a proxy for their mood changes. Therefore, mood changes, noise trading, and the optimism or pessimism of individual investors can be important factors that might help determine asset prices in capital markets.

- **Message Volume**

Message volume for a time interval t can be defined as the total number of tweets in that given time interval. Let t denote the *per diem* time interval; hence, the daily message volume for a specific stock/index on day t will be calculated as a natural logarithm of the total number of tweets per day for that particular stock/index. Given the growth of investment forums such as StockTwits, a large volume of messages is posted every day, leaving a massive amount of messages on a *per diem* basis. As with the bullishness measure, the natural logarithm transformation will also be used to control the volume of messages. The total number of tweet messages M^c is calculated as the sum of both 'buy' and 'sell' messages ($M_t^{Buy} + M_t^{Sell}$) while hold messages are ignored for the same reason outlined in the bullishness section above.

Chapter Four: Research Methodology

The following equation is used to calculate the message volume M_t^{volume} feature of StockTwits¹⁵:

$$M_t^{volume} = \ln(1 + M^c) \quad (4.27)$$

- **Level of Agreement**

Agreement among messages plays a significant role in affecting stock market behaviour. Researchers have long addressed the issue of investors' disagreement as a possible inspiration for trading (Harris and Raviv, 1993; Karpoff, 1986; Kim and Verrecchia, 1991). Following Antweiler and Frank (2004b), the level of agreement among messages is defined by calculating an "Agreement Index" as follows:

$$A_t = 1 - \sqrt{1 - B_t^2} \in [0,1] \quad (4.28)$$

where A_t is the agreement index at day t and B_t is the bullishness index at day t . The agreement index measure is commonly derived from the variance of buy vs. sell messages as a measure of the divergence between messages. The variance of B_t during time t corresponding to equation (4.23) is calculated as:

$$\sigma_t^2 = \frac{\sum_{i \in D(t)} w_i (x_i - B_t)^2}{\sum_{i \in D(t)} w_i} = \frac{\sum_i w_i x_i^2}{\sum_i w_i} - B_t^2 = 1 - B_t^2 \quad (4.29)$$

where x_i is the difference between sell or buy messages which are defined as $x_i = x_i^{Buy} - x_i^{Sell} \in \{-1, +1\}$. Again, all hold messages are excluded. w_i is the weighted message of x_i . As x_i is either -1 or +1, x_i^2 will always equal 1; this means that the following simplification would equal $\frac{\sum_i w_i x_i^2}{\sum_i w_i} = 1$ resulting in the last simplification of the variance $\sigma_t^2 = 1 - B_t^2$. The square root of $1 - B_t^2$ indicates the standard deviation of buy to sell messages.

¹⁵Particular variables such as message volume and trading volume (as will be shown later in the chapter) are calculated as $\ln(1+x)$ is calculated instead of $\ln(x)$ in order to avoid taking the log of zero when x is zero.

Chapter Four: Research Methodology

To illustrate the agreement index written in equation (4.28), it is first important to note that when all messages x_i are either bullish or bearish, that is $x_i = x_i^{Buy}$ or $x_i = x_i^{Sell}$, the agreement will equal 1 as the standard deviation of the buy to sell message represented by $\sqrt{1 - B_t^2}$ in equation (4.28) will equal 0. The agreement index (A_t) will take a value between 0 and 1. The level of agreement will be low, as the value of A_t gets closer to 0. In other words, this low value indicates high disagreement among messages. In contrast, a high agreement level will be maintained if A_t gets closer to 1. The agreement level will be zero if the numbers of sell and buy messages are equal.

B) Stock Market Indicators

The financial data have been downloaded in daily intervals for the DJIA Index from Bloomberg. Three financial variables will be considered in this research paper: Return, Trading Volume and Volatility. A brief discussion on how each of these variables is calculated will be provided in the following subsections:

- **Return**

The daily returns are calculated as the difference of the natural logarithm between the closing value of the stock price of a particular day P_t and the previous day P_{t-1} .

$$R_{it} = \ln \frac{P_{it}}{P_{it-1}} \times 100 \quad (4.30)$$

where R_{it} = Return on stock i for day t; P_{it} is the price of company i for day t; P_{it-1} is the price of company i for day t-1; and \ln = natural logarithm (the natural logarithm of the share returns was calculated to overcome any issues with non-normality in the data and is taken to constitute a non-linear transformation of the data (Brooks, 2008; Strong, 1992).

- **Trading Volume**

Trading volume is the number shares traded in a given security or an entire market in a given period of time. Trading volume is a good indicator to measure the

Chapter Four: Research Methodology

significance of the price movement of a particular security in the stock market. The higher the trading in shares in a particular period of time, the stronger the price movement (either up or down) for that period. The daily trading volume is calculated by taking the logged number of traded shares in a given day t .

$$TV_{it} = \text{Ln}(\text{number of traded share in day } t \text{ for a company } i) \quad (4.31)$$

- **Volatility**

Volatility can be defined as a measure of dispersion of returns for a given security or an index in the capital market. It measures the riskiness of an asset or an index in earning a particular rate of return. Following Garman and Klass (1980) and Alizadeh et al. (2002), the daily volatility is estimated based upon the historical opening, closing, high and low prices. They argue that volatility estimators based on historical data, namely the high, low, opening and closing prices, may contain superior information content that results in much higher efficiency than that provided by the standard volatility estimators. The daily price data (opening, closing, high and low) are obtained and used to estimate daily stock return volatility as follows:

$$\hat{\sigma}^2 = 0.511 (H_t - L_t)^2 - 0.019[(C_t - O_t)(H_t + L_t - 2O_t) - 2(H_t - O_t)(L_t - O_t)] - 0.383 (C_t - O_t)^2 \quad (4.32)$$

where $\hat{\sigma}^2$ is the variance of price change (volatility), H_t is the highest price on day t , L_t is the lowest price on day t , C_t is the closing price of day t , and O_t is the opening price of day t (all in natural logarithms). Engle and Patton (2001) and Andersen et al. (2003) have provided significant evidence that volatility is fairly long-lived. Therefore, to model this long memory series and to avoid persistence in the volatility series, changes in volatility ($\Delta v_{it} = v_{it} / v_{it-1}$) are considered in all of our corresponding analyses in this thesis rather than volatility level v_{it} .

4.9 The Econometric Model

4.9.1 The Vector Autoregressive (VAR) Model

Vector Autoregressions (VARs) are one of the most widely used classes of models in applied econometrics (Toda and Yamamoto, 1995). It is used to capture the interdependence linearity among multiple sets of time-series variables. VAR modelling is employed to estimate the intertemporal effects among all variables (k) in a VAR system where each variable has its own equation based on its own lags and the lags of the other model variables. Referring to Lutkepohl (2005), a VAR model is described as a linear function of the past values of a set of k variables (called endogenous variables) over the same sample period of time ($t=1, \dots, T$). For the i^{th} variables at a time t observation, VAR with p -th order expressed by VAR (p) is specified as follows:

$$y_t = \alpha + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \dots + \beta_p y_{t-p} + \varepsilon_t, \quad (4.33)$$

where the I -periods back observation (e.g. y_{t-1}) is called the I -th lag of y , α is the constant (intercept) with a $k \times 1$ vector, β_i is a time-invariant $k \times k$ matrix of coefficients and ε_t is a $(1 \times k)$ noise-vector (error term) conditioning that:

$$\begin{aligned} \varepsilon_t &\sim N(0, \Omega_t) \\ E(\varepsilon_t) &= 0 \end{aligned} \quad (4.34)$$

where the error term ε_t is normally distributed with Ω_t being the corresponding variance-covariance matrix and $\{\varepsilon_t = (\varepsilon_{1,t}, \dots, \varepsilon_{n,t})\}$ every error term has a zero mean (Toda and Yamamoto, 1995).

The general matrix notation of VAR (p) is expressed in the following matrix form:

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \\ \vdots \\ y_{k,t} \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_k \end{bmatrix} + \begin{bmatrix} \beta_{1,1}^1 & \beta_{1,2}^1 & \dots & \beta_{1,k}^1 \\ \beta_{2,1}^1 & \beta_{2,2}^1 & \dots & \beta_{2,k}^1 \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{k,1}^1 & \beta_{k,2}^1 & \dots & \beta_{k,k}^1 \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \\ \vdots \\ y_{k,t-1} \end{bmatrix} + \dots + \begin{bmatrix} \beta_{1,1}^p & \beta_{1,2}^p & \dots & \beta_{1,k}^p \\ \beta_{2,1}^p & \beta_{2,2}^p & \dots & \beta_{2,k}^p \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{k,1}^p & \beta_{k,2}^p & \dots & \beta_{k,k}^p \end{bmatrix} \begin{bmatrix} y_{1,t-p} \\ y_{2,t-p} \\ \vdots \\ y_{k,t-p} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \vdots \\ \varepsilon_{k,t} \end{bmatrix}$$

Writing the Equation (4.33) as one to one variable gives:

$$\begin{aligned} y_{1,t} &= \alpha_1 + \beta_{1,1}^1 y_{1,t-1} + \beta_{1,2}^1 y_{2,t-1} + \dots + \beta_{1,k}^1 y_{k,t-1} + \dots + \beta_{1,1}^p y_{1,t-p} + \beta_{1,2}^p y_{2,t-p} + \dots + \beta_{1,k}^p y_{k,t-p} + \varepsilon_{1,t} \\ y_{2,t} &= \alpha_2 + \beta_{2,1}^1 y_{1,t-1} + \beta_{2,2}^1 y_{2,t-1} + \dots + \beta_{2,k}^1 y_{k,t-1} + \dots + \beta_{2,1}^p y_{1,t-p} + \beta_{2,2}^p y_{2,t-p} + \dots + \beta_{2,k}^p y_{k,t-p} + \varepsilon_{2,t} \\ &\vdots \\ y_{k,t} &= \alpha_k + \beta_{k,1}^1 y_{1,t-1} + \beta_{k,2}^1 y_{2,t-1} + \dots + \beta_{k,k}^1 y_{k,t-1} + \dots + \beta_{k,1}^p y_{1,t-p} + \beta_{k,2}^p y_{2,t-p} + \dots + \beta_{k,k}^p y_{k,t-p} + \varepsilon_{k,t} \end{aligned} \quad (4.35)$$

Chapter Four: Research Methodology

It can be seen from the above equations that, at a time t , each variable has one VAR model regressed on its own lagged values as well as on the lagged values of each other variable where the lag length is sufficient to capture the dynamic association between the variables in the system. Determination of an appropriate lag structure is a central issue of concern in estimating the VAR model, which will be dealt with in Chapter 6. The VAR model might be augmented with some exogenous variables, namely control variables that hold constant to test the relative impact of independent variables in the regression (i.e. stock market index to control for overall market-wide effects) and/or dummy variables to account for any structural changes in the data (i.e. dummy for day of the week). There are a number of prerequisite diagnostic test procedures through which the variables under study have to pass in order for the VAR model to be implemented. Those diagnostic tests are co-integration, stationarity, autocorrelations, normality and heteroscedasticity. Appendix IV briefly explains each of these diagnostic tests and the methodology corresponding to each of them.

The VAR framework was originally developed and applied by Campbell and Shiller in 1986. It enables researchers to answer the following questions: First, can stock return be predicted from the information presented in the VAR model? Second, how volatile are stock returns to any news arriving in the market? Since this research thesis examines the forecasting power of stock micro-blogging features in predicting financial variables (return, volatility and trading volume), the VAR framework was deemed suitable as this framework seeks to establish whether there is marginally significant forecasting power in either direction (micro-blogging features have the power to predict financial variables or vice versa). As noted by Goebel et al. (2003), VAR models form a natural context based on the concept of Granger Causality by which the directed influence can be measured. Unlike the empirical method followed initially by Fama and French (1989), who used long-horizon regressions to test for asset return predictability by regressing asset return at increased time horizons (Campbell and Shiller, 1988 a and b), short-horizon vector autoregression (VARs) is used to estimate the intertemporal correlations between stock micro-blogging features and the stock market. This method has several advantages. First, it avoids the small sample biases inherent in long-horizon regressions. Second, it prevents overlapping

Chapter Four: Research Methodology

regressions (Hodrick, 1992). Third, it allows for feedback and interactions between all model variables in the form of lead-lag relationships.

4.9.2 Quantile Regression Approach

The quantile regression (QR) model, as first proposed by Koenker and Bassett (1978), provides estimates of linear relationships between the regressors over specified quantiles of the regressand. It offers a new approach to estimating the conditional quantiles of a dependent variable y , given one or more explanatory variables. Unlike the estimated coefficients produced by the traditional Ordinary Least Square (OLS), the coefficients estimated using QR are more efficient and robust since the QR model has focused mainly on the location model, and the effects of conditioning are restricted to a location shift. The quantile regression framework is used to examine the influence of a lagged change in bullishness on all quantiles of the current return¹⁶.

One of the main attractions of employing quantile regression in this research is the fact that returns are not normally distributed, as one would expect different effects of predictor variables over various quantiles of returns distributions. This study seeks to investigate whether or not the impact of lagged change bullishness is different across quantiles of contemporaneous return. The purpose of the study is to model the quantile of stock return for a given bullishness level based on a linear model as well as considering the asymmetric non-linear behaviour of investor sentiments (bullish and bearish sentiments) in stock return. In many cases, quantile regression estimates are quite different from OLS models. These results carry crucial implications for the linkage between investor sentiment and stock markets.

In this thesis, the relationship between investor sentiment and stock returns is revisited by using the QR technique, developed by Koenker and Bassett (1978). The following conditional quantile model is estimated as follows:

¹⁶Quantile regression has been widely used in many areas of empirical finance and applied econometrics. Feng et al. (2008) and Ma and Pohlman (2008) investigate the performance of momentum portfolios based on quantiles momentum measures of past performance based on quantiles of past returns. Chuang et al. (2009) investigate the dynamic relationship between stock return and trading volume based on quantile regression and find evidence of a causal effect of lagged volume on return of opposite signs at lower and upper quantiles but not central quantiles. Baur et al. (2012) employ a quantile regression approach to examine the predictability of return across a range of quantiles of the conditional return distribution and find that the autoregressive model follows a decreasing pattern (positive (negative) dependence on past return at lower (upper) quantiles of the conditional return distribution). Alagidede and Panagiotidis (2012) investigate the short-term relationship between stock return and inflation rate for the G7 countries, and a positive relationship was found in most countries under study, with the magnitude of this relations increasing as it moved toward upper quantiles.

Chapter Four: Research Methodology

$$R_{it} = \alpha(\tau) + \beta(\tau)\Delta\bar{B}_{it}^{**} + \gamma_1(\tau)MKT_t + \gamma_2(\tau)NWK_t + \varepsilon_{it}(\tau) = x'_{it}\boldsymbol{\theta}(\tau) + \varepsilon_{it}(\tau), \quad (4.36)$$

where τ denotes the τ -th conditional quantile of stock i 's return, $\boldsymbol{\theta}(\tau)=[\alpha(\tau), \beta(\tau), \gamma_1(\tau), \gamma_2(\tau)]'$, $\Delta\bar{B}_{it}^{**}$ is the shift in bullishness of the corresponding stocks, MKT and NWK are the market control variables and the first day of the week dummy respectively, as defined earlier. The estimated coefficient of $\beta(\tau)$ is our main concern in this model specification, which can be interpreted as a parameter estimate of a specific τ -th conditional quantile.

Moreover, to uncover the asymmetric effect of sentiment on stock returns using the QR, Eq. (4.36) is re-specified by including the lagged bullish and bearish sentiment effects separately in the model as follows:

$$R_{it} = \alpha(\tau) + \beta_1(\tau)\Delta\bar{B}_{it}^{**}D_{it} + \beta_2(\tau)\Delta\bar{B}_{it}^{**}(1 - D_{it}) + \gamma_1(\tau)MKT_t + \gamma_2(\tau)NWK_t + \varepsilon_{it}(\tau) = z'_{it}\boldsymbol{\psi}(\tau) + \varepsilon_{it}(\tau), \quad (4.37)$$

where $\boldsymbol{\psi}(\tau)=[\alpha(\tau), \beta_1(\tau), \beta_2(\tau), \gamma_1(\tau), \gamma_2(\tau)]'$. The aim of this model specification is to assess the influence of both bullish and bearish shifts in sentiment on the different conditional quantiles of stock returns measured by $\beta_i(\tau)$, $i = 1$ and 2 .

The parameter vectors $\boldsymbol{\theta}(\tau)$ and $\boldsymbol{\psi}(\tau)$ are estimated using linear programming techniques (see Koenker and D'Orey, 1987)¹⁷ by solving the following minimisation problems:

$$\min_{\boldsymbol{\theta}} \sum_{t=1}^T (\tau - \mathbf{1}_{\{R_{it} < x'_{it}\boldsymbol{\theta}\}}) |R_{it} - x'_{it}\boldsymbol{\theta}|, \quad (4.38)$$

$$\min_{\boldsymbol{\psi}} \sum_{t=1}^T (\tau - \mathbf{1}_{\{R_{it} < z'_{it}\boldsymbol{\psi}\}}) |R_{it} - z'_{it}\boldsymbol{\psi}|. \quad (4.39)$$

¹⁷For more details on the QR techniques, the reader is directed to the surveys by Buchinsky (1998) and Koenker and Hallock (2001).

Chapter Four: Research Methodology

Both Eqs (4.36) and (3.37) are estimated with 9 quantiles (i.e., $\tau = 0.05, 0.1, 0.25 \dots 0.95$). The entire distribution of the regressor conditional on the regressand is traced as τ increased from 0 to 1. (More details of these analyses will be provided in Chapter Seven).

Summing up, quantile regression provides a holistic picture of the relationship between two variables at different points of a conditional distribution of the dependent variables. It therefore provides a promising insight into the way of describing the whole distribution while adding value in explaining the relationship between the regressors and the independent variable, which may evolve across its conditional distribution.

4.10 Chapter Summary

This chapter offered an outline of different methodological approaches that have been utilised within the information systems and finance fields where the most appropriate approaches were selected for guiding and presenting this particular research. This chapter has primarily emphasised the selection of the most effective approaches to capturing and preserving the depth and richness of the data throughout the research process.

An overview of the two research paradigms (positivist and interpretive) that exist in the domain of IS and finance research was provided in order to demonstrate that the positivist stance should be the philosophical foundation of this research. Adopting the positivist stance allows the researcher to measure and observe the attitudes and behaviours of individuals. Moreover, it enables the researcher to relate the facts and causes of social phenomena while remaining detached and independent from what is being observed. Following this, a general discussion of quantitative and qualitative research approaches was provided in this chapter, justifying the reasons for the adoption of the quantitative approach as the more appropriate approach. The quantitative approach enables the researcher to empirically test the research theories and validate and understand the conceptual framework. The data collection method that was employed included the use of secondary data, which reduces the likelihood of researcher bias that may affect the validity of the research. Textual analysis techniques alongside financial econometrics modelling were judged to be the appropriate means of analysing the data.

Chapter Four: Research Methodology

As this chapter has presented and justified the positivist and quantitative approaches as those most suitable for this study, this has now set the stage for presenting, reporting and discussing the key findings and results of the empirical tests from textual analysis and various statistical and econometrics modelling techniques.

CHAPTER FIVE: TEXT MINING ANALYSIS AND FINDINGS

5.1 Introduction

The previous chapter explained and justified the research methodology adopted for this study. The intention of this chapter is to present the general findings of data analysis of StockTwits while providing a comprehensive discussion from the analysis of the manual and automated classifications of StockTwits data using different machine-learning algorithms. A comparative analysis of three different machine-learning algorithms is performed in this chapter in order to determine the most suitable and accurate techniques for sentiment analysis of StockTwits messages. This chapter also presents the findings on the effectiveness of feature selection in improving sentiment classification accuracy of different classifiers. Additionally, this chapter provides two different applications of feature selection, applying both filter and wrapper approaches.

This chapter consists of ten sections including this introduction. Section 5.2 presents the findings of the manual classification process of StockTwits postings. Section 5.3 provides a comparative analysis of the automated classifications of three different machine-learning processes using different performance evaluation methods. Section 5.4 investigates the effectiveness of the feature selection methods in improving the sentiment classification performance of all studied classifiers while highlighting the extent to which each of these classifiers benefits from performing both the filter and wrapper approach. The selection of the best classifier based on the general automated classifications performance and in accordance with its effectiveness in performing feature selection approaches is discussed in section 5.5. Section 5.6 presents the experimental results of training and testing to investigate the significance of the ability of the selected classifier to classify any instance deployed from the StockTwits population sample used in this study. Section 5.7 summarises the results of the overall classification distributions of all tweet messages per sentiment class. Two different applications of feature selection methods using two different machine-learning classifiers - Bayesian Network with a wrapper approach and Decision Tree with a filter approach - are implemented in Sections 5.8 and 5.9 respectively. Finally, Section 5.10 offers a brief summary of this chapter.

5.2 StockTwits Sentiment Hand-Labeling

In order to manage the huge amount of StockTwits messages collected for this study, a random sample of tweet postings is selected and manually classified; this will be used as training set for different machine-learning models. The tweets are labelled buy (1), hold (2) and sell (3). The results of the percentage allocation of the manual classifications of tweet messages into the three distinct classes are shown in Table 5.1.

Table 5.1: The manual classifications of StockTwits messages				
Class	Buy	Hold	Sell	Total
Numbers	1,361	590	941	2,892
Percentage	47.06%	20.40%	32.54%	100%

As can be seen from the table above, roughly half of these messages were considered to be “buy” signals (47.06%). The remaining messages are for “sell” signals, which are roughly three quarters of “buy” signals as (32.54%) and for “hold” signals as (20.40%). The results of this study indicate that this stock micro-blogging forum seems to be more balanced in terms of the distributions of buy vs. sell messages than internet message boards where the ratio of buy vs. sell signals appears to be unbalanced, ranging from 7:1 Dewally (2003) to 5:1 Antweiler and Frank (2004b). As the “hold” messages formed a relatively small percentage of 20.40%, this finding does not support the previous study by Sprenger and Welpe (2010), who found that almost half of the messages manually classified were considered to be “hold” signals. The findings of the current study concerning the tiny proportion of “hold” signals indicate that little noise is involved in the StockTwits forum about the DJIA index, which may also imply the limited effect of noise traders’ activities in this forum. On the other hand, the higher distribution of buy and sell messages may provide evidence that there is more relevant financial information present in such forums. More excitingly, the greater proportion of buy messages may serve as a proxy for positive investor sentiment expecting stock prices to rise as investors are more bullish and optimistic and are therefore demanding more of those stocks in their portfolios. To understand the nature of classified messages, it is helpful to look at examples. Table 5.2 provides a few typical examples of manually classified tweets from the training set including the manual coding.

Table 5.2: Sample tweets from training set with manual classification

Sample Tweets (Training Set)	Manual Classification
"Our highest long as of today low \$JPM and \$BAC.//LOL"	Buy
"\$xom \$intc \$dvn \$ko \$cm \$ftse some analysis on these charts"	Hold
"Short \$NKE http://chart.ly/jmbomde "	Sell
"\$KO http://stks.co/3OvK Breaks yesterdays high will add! Bullish"	Buy
"\$CAT In again for giggles at 81.16... 2 Dec 80 Put for DCA of \$2.32 Average entry at 81.25"	Sell
"\$SBUX The Starbucks Trade http://stks.co/nDQ6 \$DNKN \$MCD \$ARCO \$GMCR"	Hold
"\$T good stock for buying... http://stks.co/t04i "	Buy
"\$GS we are buyers on dips. (Shares and long term calls)"	Buy
"\$NKE down over 2% now. Making new lows trying to break \$97"	Sell
"\$CAT Looks ugly down there http://chart.ly/gk4hbm8 "	Sell

Looking at the most common words associated with each class, it is obvious that some general features occur very frequently in all three classes (e.g. figures and ticker names and external links). However, beyond these universal features, there is a unique pattern that reasonably distinguishes the linguistic bullishness of each of the three classes. For example, positive words such as “good” and “high” are the most common words likely to be found in buy messages in addition to financial words such as “buy”, “long” and “call”, which in the financial context give a clear sign that investors are expecting a particular stock to rise. In contrast, the most common words likely to appear in sell messages are negative words such as “down”, “ugly”, “break” and “low”, as well as words such as “sell”, “put”, “loss” and “short” which give a clear signal that users are expecting the discussed stock to fall. These results match those observed in earlier studies by Sprenger et al. (2014) and Tetlock et al. (2008). However, if the tweet message contains external links to long articles or charts about the stocks, in which more neutral words appear, such as the product name (e.g. “Aircraft”, “BigMac”, “Window7”), it is generally labelled hold. Therefore, in hold messages the positive and negative words are much more balanced and neutral words dominate.

5.3 Model Building in Weka

Several types of models have been made available in Weka, each with different algorithms to build a model. The most commonly used machine-learning algorithms include Bayesian Networks, Decision Trees, Neural Networks, Fuzzy

Chapter Five: Text Mining Analysis and Findings

Networks, Support Vector Machines, Genetic Algorithms and many more. However, to keep the scope of this research more focused, Bayesian Networks (Naïve Bayes (NB)), Decision Trees (Random Forest (RandF)) and Support Vector Machines (Sequential Minimal Optimisation (SOM)) are used to perform the text analysis tasks. The performances of these three models were then evaluated on the training data and compared in order to select the best model. The following subsections describe the testing of the three models using two different methodologies: testing on the training set and testing by tenfold cross-validation.

The input for all the models used in this study comes from a training corpus of 2,892 tweet messages. Ideally, the model should have been trained on more data instances as it is expected that the accuracy of the models will increase when larger training datasets are handled. However, there is always a trade-off between high model accuracy and the risk of model over-fitting. Therefore, the training data should be kept small to avoid the risk of over-fitting associated with large amounts of training data.

There are two different methodologies normally used in evaluating machine-learning performance: testing on training sets and testing by tenfold cross-validation. However, before starting to test and analyse the models, it is necessary to highlight some important facts about the Weka machine-learning environment. First, testing on the training data will always show better results and will be optimistic compared with what might be expected from the stratified tenfold cross-validation, while the latter always provides a more conservative measure of classification accuracy. In addition, k-fold cross-validation provides the best generalisability and helps overcome the risk of model over-fitting as each of these folds check whether the learned model over-fits on the validation set. The stratified cross-validation provides a more realistic picture than testing on the full training sets. Therefore, it is worth mentioning that our main focus in the analyses will be on the results of the tenfold cross-validation (Whitten and Frank, 2011) in order to strengthen the validity of the results while providing only a small window to briefly discuss the results of testing using training data. In line with the standard metrics of Information Retrieval, recall, precision and F-measure (Whitten and Frank, 2011) are the reported measures used to evaluate the performance of the predictive model.

Chapter Five: Text Mining Analysis and Findings

The following table (Table 5.3) presents a consolidated summary of all the performance metrics of the three classifiers using tenfold cross-validation. Although it is evident that there is no clear winning classifier in terms of the performance evaluation method used, the Random Forest Classifier is possibly the best classifier in terms of almost all the metrics, as shown in the table below.

Table 5.3: Summary results of the classification performance evaluation of NB, RandF and SMO

Weighted Average Metrics for (buy, hold and sell) class	Classifier		
	Naive Bayes	Random Forest	SMO
Accuracy Rate	62.80%	66.70%	65.25%
Correctly Classifies Instances	1,815	1,929	1,887
Incorrectly Classified Instances	1,077	963	1,055
TP Rate	62.80%	66.70%	65.20%
FP Rate	21.80%	20.80%	24.60%
Precision	62.90%	66.50%	65.90%
Recall	62.80%	66.70%	65.20%
F-Measure	62.60%	66.20%	64.00%
ROC Area	77.60%	79.80%	73.60%

As it can be seen from Table 5.3, the tenfold cross-validation experiments achieved accuracy figures of 66.70%, 62.80% and 65.20% where 1,929, 1,815 and 1,887 instances were correctly classified out of 2,892 for RandF, NB and SMO respectively. The numbers reported in Table 5.3 clearly show that the Random Forest decision tree classifier outperforms the Naive Bayes and SMO classifiers in predicting the investor sentiment class (buy, hold and sell) of StockTwits postings. The weighted averages of the three classes of RandF classifier are also shown in the table at 66.50%, 66.70% and 66.20% for precision, recall and F-measures respectively. More elaborative details of the performance analysis of all three classifiers undertaken in this study will be provided in Appendix V.

The following chart shows the graphical representation of the comparative performance of the three discussed classifiers using some of the important measures given in Table 5.3. As it can be seen in Figure 5.1, all classifiers perform more or less the same, while Random Forest shows a slightly better performance than Naive Bayes and SMO.

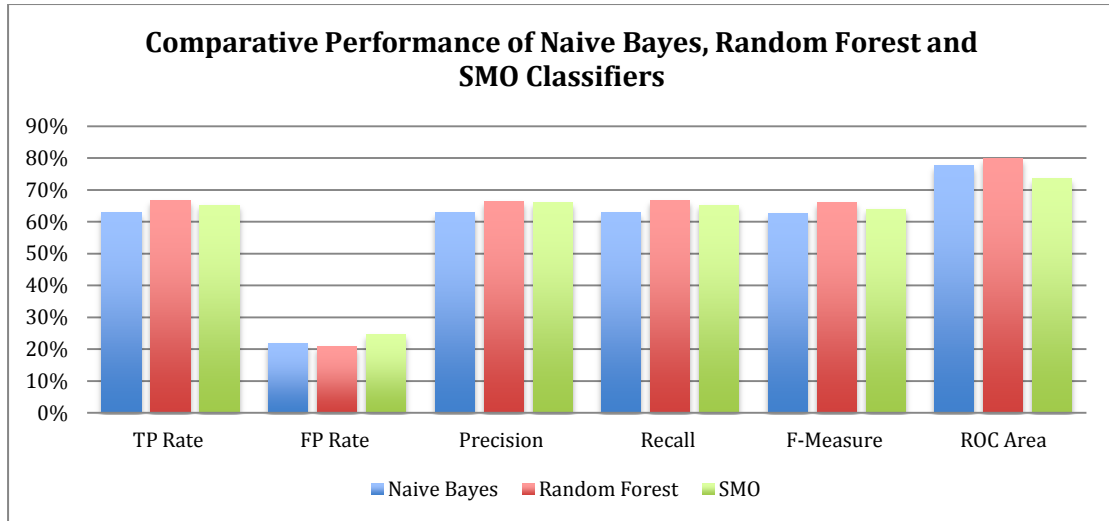


Figure 5.1: Comparative Performance of NB, RandF and SMO classifiers

5.4 Feature Selection

Feature selection (FS), an essential pre-processing step to machine learning, is effective in handling data-mining tasks, omitting irrelevant data, reducing dimensionality and improving prediction accuracy of the classifiers. Filter and wrapper are the two most commonly used methods to perform feature selection. The effects of the feature selection on the quality of different classifiers employed in this study were measured by the overall classification accuracy, which is used most frequently in machine learning.

5.4.1 Filter Approach

To extract the filter subset, a ranker search method (Mark et al., 2009) was used in conjunction with the information gain criteria where the worth of an attribute is evaluated by measuring its information gain (IG) score with respect to the class. Table 5.4 shows the result of filter feature selection with the listed terms ranked according to their IG values.

As it can be seen from the Table 5.4, 47 terms (including the “sentiment” class) are retained after performing the filter approach using information gain criteria. The terms listed in the table are ranked according to their relevancies; those at the beginning of the list (indicated by the serial number) are most relevant and the relevancy decreases as one goes down the list. The information gain (IG) value is

Chapter Five: Text Mining Analysis and Findings

reported next to each term. For example, the terms ‘ID’ and ‘short’ appear to be the most significant of all the listed terms, with IG values of 0.0902 and 0.0706, while ‘run’ is the least important term, with an IG value of 0.0034.

Table 5.4: Features selected under filter approach using information gain criteria

Sr.	Feature	IG	Sr.	Feature	IG
1	ID	0.0902	24	mrk	0.0072
2	short	0.0706	25	utx	0.0070
3	cat	0.0320	26	move	0.0070
4	csc	0.0251	27	stop	0.0066
5	bearish	0.0225	28	bull	0.0063
6	aapl	0.0205	29	unh	0.0059
7	bullish	0.0196	30	volum	0.0056
8	cvx	0.0170	31	pfe	0.0051
9	nice	0.0159	32	target	0.0047
10	breakout	0.0142	33	support	0.0046
11	lower	0.0141	34	msft	0.0044
12	xom	0.0123	35	bounc	0.0044
13	break	0.0121	36	entri	0.0043
14	look	0.0115	37	sell	0.0042
15	strong	0.0114	38	set	0.0042
16	quot	0.0099	39	weak	0.0042
17	current	0.0096	40	gap	0.0041
18	high	0.0089	41	head	0.0041
19	buy	0.0087	42	market	0.0040
20	goog	0.0082	43	flag	0.0039
21	post	0.0080	44	bottom	0.0038
22	report	0.0079	45	bought	0.0037
23	spi	0.0078	46	run	0.0034

Note that the feature “ID” is shown at the top of the list as each post indicated by the user ID in the “term document matrix” of the training set.

• Classification Performance with the n “Best Ranked” InfoGain

Experiments were performed using the three machine-learning models adopted in this thesis: Naive Bayes (NB) Algorithm, Random Forest (RandF) classifiers, and sequential minimal optimisation algorithm (SMO). Using the Information Gain method, the classification performance is measured for the subsets consisting of the n “best ranked” features (47 attributes including the sentiment class) as reported in Table 5.4. The classification experiments are then repeated, each time with a certain percentage reduction of the feature sets where the features towards the bottom of the list with the lowest information gain will be removed first while retaining the most (best) relevant features that will be used as inputs to machine-learning models. Table 5.5 shows the average classification accuracy for the information gain subsets for the

Chapter Five: Text Mining Analysis and Findings

three machine-learning models (NB algorithm, RandF classifier and SMO) over various numbers of subset reductions of the n “best ranked” features. The classification accuracy results of the reduced features subsets are shown separately for the three classifiers’ algorithms below.

Table 5.5: The Best Overall Classification Accuracy (in %) for the Information Gain subsets for NB, RandF and SMO.			
Attributes	NB	RandF	SMO
All Attributes 100	<i>62.80</i>	<i>66.70</i>	<i>65.25</i>
n"Best ranked" 47	65.42	63.90	62.55
44	65.35	64.56	62.14
43 (NB best n wrapper)	65.18	63.90	61.70
42 (RandF best n wrapper)	65.18	64.35	61.76
38	65.20	63.55	60.96
35	65.18	64.11	60.93
32	65.32	63.43	60.51
29	65.14	63.42	60.51
26	64.97	62.72	60.10
23	64.18	62.00	60.37
20	63.38	62.00	59.92
17	63.24	62.03	59.65
14	63.03	61.93	59.47
11	61.96	60.48	58.92
8	61.38	58.99	58.78
5	59.82	56.40	56.36

The average classification accuracy for each learning algorithm using all features (100 attributes) are highlighted in italics while the best average results achieved over the feature reduction (filter) methods are highlighted in bold. As shown in Table 5.5, the Naive Bayes classifier performed well when the number of features was reduced to 47 (about 50% removal) by IG. A comparison of the three machine-learning classifiers reveals that the highest accuracy level using information gain subsets is achieved by the Naive Bayes algorithm, followed by Random Forest classifiers and then Support Vector Machine (SMO), with average accuracies of 65.42%, 64.56% and 62.55% respectively.

The results shown in Table 5.5 reveal that SVM (SMO) did not benefit from filter feature selection. This study’s results corroborate the findings of a great deal of previous work reported in text classification (Yang and Pedersen, 1997; Rogati and Yang, 2002; Brank et al., 2002; Liu, 2004). Compared to NB classifiers, SVM (SMO) achieves the best average performance accuracy when all the features were given to SMO. This finding supports previous research into this field of study, in which SVM tends to outperform other classifiers’ algorithms and results in the highest

Chapter Five: Text Mining Analysis and Findings

classification accuracy when all features are included in the classification experiments (Taira and Haruno, 1999). All the algorithms are performing more or less the same; nevertheless, when looking simultaneously at the size of the subset features and the highest accuracy, it can be seen that Random Forest achieves the highest accuracy with smaller subsets (of 44 attributes) compared to Naive Bayes and SMO, both of which achieved their highest accuracy with 47 attributes. Figure 5.2 shows the behaviour of the three machine-learning classifiers over different subset feature reductions using the information gain method.

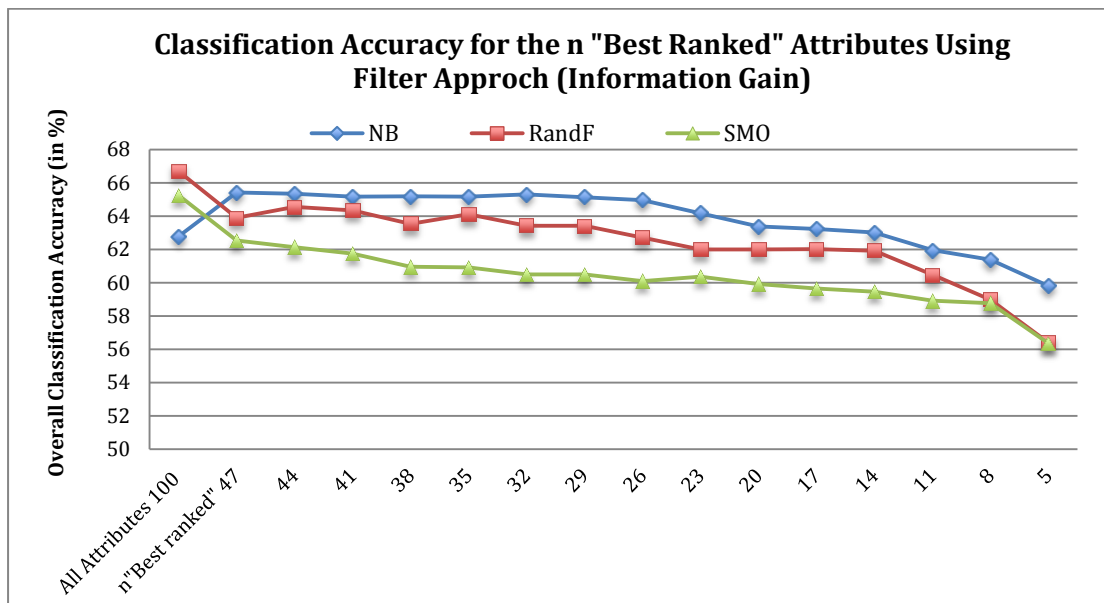


Figure 5.2: The overall classification accuracy for the “best ranked” attributes by IG criteria.

As the Figure 5.2 shows, the classifiers show marginally different behaviour over different subset reductions. For example, the classification accuracy is quite stable for the Naïve Bayes and Support Vector Machine (SMO) but it tends to fluctuate for the Random Forest classifier. All machine-learning methods show a degradation in the classification accuracy when the number of features in the IG subsets is reduced to 17 features or less, where the overall classification accuracy tends to decrease proportionally with the reduction of features. In general, performing the filter approach based on Information Gain evaluators improves the performance of the Naive Bayes classifier, as it performs slightly better than when all feature sets are involved in the classification problem where the average accuracy level increased from 62.8% to 65.42%. This finding is in agreement with that of Liu (2004), who

Chapter Five: Text Mining Analysis and Findings

showed that the classification performance of Naive Bayes classifiers improved significantly when using feature selection based on Information Gain (IG). Meanwhile, the results of the experiments indicated that Random Forest and SMO did not benefit from filter methods, as the classification accuracies achieved by both classifiers are only slightly worse than the accuracy achieved with the complete feature subsets.

5.4.2 Wrapper Approach

Since three different machine learning classifiers (NB, RandF and SMO) are applied in this thesis, Table 5.6 presents the optimum feature attributes selected under the wrapper method for each classifier independently along with their average classification accuracy. As it can be seen from Table 5.6, performing the wrapper feature selection reveals that 43, 42 and 66 attributes are the best attribute combinations for NB, RandF and SMO classifier respectively. It is worth noting that there was a great variation in the features subset chosen under the wrapper method for the three classifiers. This tends to support the findings of Kohavi (1995) who demonstrates that the feature set is best considered part of the classifier algorithm chosen and that selected features are tailored to a particular algorithm used in the attribute evaluator. As he strongly argues, it is unlikely that a set of features selected will be the optimum for all classifiers.

Table 5.6: The Average Classification Accuracy of the “n” Best Attributes Selected Under Wrapper Method for NB, RandF and SMO Classifiers		
Classifiers	Wrapper Attributes	Classification Accuracy (in %)
NB	43 Attr.	66.42
RandF	42 Attr.	68.08
SMO	66 Attr.	66.94

From the attributes selected under the wrapper method that have been analysed in Table 5.6, the results of classification performance using the final set of features for each of the three classifiers are listed in the third column. Interestingly, all three machine-learning classifiers show a good performance under the wrapper method. This indicates that the wrapper approach, as a feature selection method, resulted in

Chapter Five: Text Mining Analysis and Findings

statistically significant improvements in classification performance over the use of the full feature set for NB, RandF and SMO classifier. From the results shown in Table 5.6, Naive Bayes achieves an accuracy of 66.42% with 43 features, while SMO achieves an accuracy of 66.94% with 66 attributes that are regarded as the optimum features to maximise classification performance. However, Random Forest attains a maximum accuracy level of 68.08% with only 42 attributes compared to NB and SMO.

5.4.3 Comparative performance of Classifiers' Algorithms under Filter and Wrapper Methods

It is very interesting at this point to consider a comparison of the two Feature Selection methods (FS) for all learning algorithms of interest in this research study. Table 5.7 shows a comparison of the accuracy achieved with both FS methods, filter (IG) and wrapper, for NB, RandF and SMO classifier.

Classifiers	NB	RandF	SMO
Methods/Selected Attributes	43 Attributes	42 Attributes	65 Attributes
Wrapper	66.42%	68.08%	66.94%
Filter (InfoGain)	65.18%	64.35%	NA*

* Information Gain (IG) returns only 47 out of a total of 100 attributes.

As Table 5.7 shows, the wrapper subsets achieve a better accuracy level compared to IG results (see Table 5.5). These results may be due to the fact that wrapper methods have the ability to discover small subsets of the most accurate features from StockTwits datasets that can better predict the three distinct classes (buy, sell or hold). Another possible explanation is the predictive power of the wrapper methods in selecting the most accurate feature subsets compared to the filter methods, as previously outlined by many researchers (Huang et al., 2008; Li and Guo, 2008). Moreover, when looking simultaneously at both the size of the attributes and the accuracy level, it can be seen that wrapper subsets achieve the best accuracy with smaller features. For example, NB achieves a maximum accuracy of 65.42% with 47

Chapter Five: Text Mining Analysis and Findings

attributes while it achieves a better accuracy of 66.42% with only 43 attributes under the wrapper method.

Comparing wrapper results with the IG results, considering the same number of attributes (refer to Table 5.7), it can be seen that the wrapper approach clearly outperforms IG. All three algorithms show greater improvements in accuracy when the wrapper is employed for feature selection rather than the filter (IG). The improvement varies from a slight improvement, in NB (from 65.18% with filter (IG) to 66.42% with wrapper), to a statistically significant improvement, as with the RandF classifier where the classification accuracy jumps to 68.08% using the wrapper method. SMO classifiers show a slight improvement in performance with the wrapper method compared to when all subsets are involved. Generally, the wrapper subset achieves better results and leads to better accuracy than IG subsets of comparable sizes. Therefore, the wrapper approach is typically regarded as superior to the filter approach in finding the most accurate feature subsets. However, the wrapper tends to be computationally more expensive than the other feature reduction methods. Therefore, research in the feature selection field is still investigating this problem and is attempting to overcome this disadvantage of the wrapper.

Having discussed the analyses of the findings and results of different experiments with the three learning algorithms (NB, RandF, and SMO) in Weka, it is important to decide which machine-learning classifiers are best suited to the classification problem of StockTwits data for this research thesis.

5.5 Selecting the Best Algorithms

The overall classification accuracy of the three learning algorithms is good in general. However, when comparing the performance of these machine-learning methods, it can be seen that the Random Forest (RandF) classifier always achieves the best results both when the complete feature set is used and when performing feature selections (wrapper and filter methods). The SMO classifier seems to be sensitive to the size of feature subsets. For the average filter (IG) results (Table 5.5 and Figure 5.2), SMO shows the lowest classification accuracy; meanwhile, with the wrapper approach, SMO results improve as more subset features are selected. On the other hand, the Naive Bayes Classifier tends to benefit from both feature selection methods,

Chapter Five: Text Mining Analysis and Findings

as its classification accuracies under filter and wrapper are better than when all features are used.

To sum up, while our findings reveal that the decision tree (Random Forest) is the best classifier, as indicated by the highest performance accuracy in all experimental results, followed by the Naive Bayes algorithm, both classifiers have the added benefit of visualising the relationships (by building the J48 Decision Tree Model and Bayesian Networks) between the selected features for this classification problem (sentiment prediction). Consequently, a better understanding of the relationships among the most relevant features will enable us to make a better prediction of features belonging to each of the three classes (buy, sell and hold). Later in this chapter two applications of Decision Tree models and Bayesian Networks will be discussed in relation to the prediction of stock micro-blogging sentiments. However, before proceeding with this, another important task in data mining must be first performed. This is the training and testing task, which will be provided in the next section.

5.6 Training and Testing

This section presents the experimental results of training and testing based on the two methods discussed previously in Chapter 4. Since the Random Forest classifier has proved best in classifying StockTwits data for this research study, it will be used to run the repeated training and testing experiments

5.6.1 Method (1): Training and Testing Using Automatic Percentage Split (66% Training and 33% Testing) Using One Dataset

Table 5.8 shows the results of the classification accuracy of Random Forest running each time with different random-number seeds. The mean is the total success rate or accuracy rate divided by the total number of experiments, which in this case is 10, i.e. 63.15 %; this is considered a more reliable estimate than the accuracy rate obtained when one random seed is chosen. The variance can be calculated by taking the standard deviation from the mean, subtracting the mean from each of the values, squaring the results, adding them up and then dividing by $n-1$, i.e. 2.79%. Taking the square root of the variance (2.79%) results in a standard deviation of 1.67%. From the

Chapter Five: Text Mining Analysis and Findings

results shown in Table 5.8, it can be seen that the real performance of Random Forest classifiers on the StockTwits dataset is approximately 63% plus or minus approximately 3% of the mean variance. It can be said that the accuracy of the Random Forest is anywhere between 60% and 66%.

Number of seeds	% Accuracy	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
1	65.66	2.51	6.32
2	62.65	-0.50	0.25
3	63.86	0.71	0.51
4	62.65	-0.50	0.25
5	63.1	-0.05	0.00
6	62.5	-0.65	0.42
7	63.7	0.55	0.31
8	64.76	1.61	2.60
9	59.34	-3.81	14.49
10	63.25	0.10	0.01
Total $\sum x_i$	631.47		25.15
Mean \bar{x}	63.147	Variance	2.79
		Standard Deviation	1.67

5.6.2 Method (2): Training and Testing Using Supplied Test Set of Two Separate Datasets (Training set (In-Sample Set) and Testing Set (Hold-Out Set))

Using the supplied test set by training on the entire first ten months' corpus (from April 2012-January 2013) while testing on the remaining two months' corpus (February 2013-March 2013) yielded an accuracy of 60.50%. The Weka output results are shown in Table 5.9 below.

Class	True Positive	False Positives	Precision	Recall	F-Measure	ROC Area
Buy	63.31%	32.0%	69.10%	63.10%	66.00%	68.30%
Hold	36.90%	4.60%	56.50%	36.90%	44.70%	78.00%
Sell	66.10%	30.70%	51.50%	66.10%	57.90%	71.00%
Weighted Average	60.50%	27.80%	61.50%	60.50%	60.04%	70.50%

As it can be seen from the findings of the two methods of training and testing, both the automatic % split using one dataset and the supplied test set using two

Chapter Five: Text Mining Analysis and Findings

separate training and testing sets consistently yielded accuracy levels somewhere between 60% and 66% accuracy. Since the supplied test set (in-sample and hold-out set) methods are considered more reliable than randomly split datasets, the accuracy rate achieved by the supplied test set of 9i8imii60.50% - which is still in the range of the accuracy interval of the percentage split (60-66%) - will therefore be used to apply the classification results to the entire population of StockTwits data.

5.7 Overall Classification Distribution

From the confusion matrix of the output results of the supplied test set, Table 5.10 provides a comparison of the manual classification of hold-out messages and the automated classification of the Random Forest algorithm.

Table 5.10: Random Forest Classification Accuracy of Supplied Test Set and the overall classification distribution				
Classified by Algorithm				
Class	Buy	Hold	Sell	Manual Classification
Buy	315	26	158	499
Hold	47	48	35	130
Sell	94	11	205	310
Total classified by Algorithm	456	85	398	939
% Classification by Algorithm As Per Class	48.56%	9.05%	42.39%	100%

Table 5.10 provides the buy-hold-sell matrix entries of the hold-out sample (939 messages) and the prediction accuracy of the classification algorithm with respect to the training (in-sample set) of 1,953 messages. The total rows of Table 5.10 show the actual share of 939 hand-coded messages that were classified as buy, hold or sell whereas the total columns represent the share of the messages that were automatically classified as per class by the algorithm. The last row line provides summary statistics of the percentage distribution of the out-of-sample classification of each class that will then be deployed and aggregated for the daily ticker level analysis. The results of Table 5.10 suggest that the algorithm performs reasonably well, as indicated by the relatively small numbers of misclassifications in each sentiment class.

Table 5.11 shows the assign labels for the entire set of StockTwits postings producing results of buy (140,350), hold (26,157) or sell (122,517) postings. As with Antweiler and Frank (2004b), the hold postings are removed from the analysis as they

Chapter Five: Text Mining Analysis and Findings

are considered noise and convey neutral opinions, while only the postings with relative sentiments ($140,350+122,517=262,867$) remain useful and relevant for further analysis.

Class	Manual Classification (in %)	Automatic Classification (in %)	Total Tweets per class
Buy	47.60	48.56	140,350
Hold	20.40	9.05	26,157
Sell	32.54	42.39	122,517
Total	100%	100%	289,024

Based on the weight assigned to StockTwits messages, the distribution of the postings as buy, sell and hold classes reveals that the highest percentage is devoted to the “buy” message. This finding is in agreement with Dewally (2003), who found that most of the messages in online investment often represent “buy” signals.

In the next two sections, two applications of feature selection (filter and wrapper approaches based on the Decision Tree model and Bayesian Network respectively) will be discussed more extensively. The aim is to clearly demonstrate the interactions of the selected features under filter and wrapper methods that provide better predictions of investor sentiments (buy, hold or sell) using StockTwits postings. The novelty of these two applications lies in the approach adopted, where text-mining tasks are combined with feature selection methods and machine-learning algorithms to predict an intelligent trading support mechanism that will help investors to make profitable investment decisions concerning a particular security in the capital market. In these two applications, both Bayesian Network model and Decision Tree model were adopted since both have the advantage of visualising relationships between selected features, which makes them the most suitable techniques for sentiment prediction in the stock market.

5.8 Application (1): Application of Wrapper Approach: Bayesian Network Model for Prediction of Investor Sentiment in Capital Market

This research study takes a different approach by integrating text-mining techniques, the wrapper approach and a Bayesian Network model to extract relevant features from StockTwits data to predict trading decisions (buy/hold/sell). The aim is

Chapter Five: Text Mining Analysis and Findings

to investigate the interactions between the selected features and their ability to predict investors' sentiments quarterly over different periods of the year. The transparency and visibility of the connected relationships between nodes and parents in the Bayesian Networks model makes it a more suitable approach for feature selection and prediction of sentiments in the stock market.

5.8.1 Experiments and Analysis

The experiment aims to predict investors' sentiments regarding a particular StockTwit post of DJIA companies on whether to buy, hold, or sell. The one-year training data are split into four subsets, each of which represents a quarter of the year's data. A prediction model is built for each subset using four different machine-learning algorithms: Bayes Net, Naive Bayes, Random Forest and Sequential Minimum Optimal (SMO). The performance is used to evaluate the efficiency of each of these classifiers based on wrapper feature selection. A textual visualization tool called Wordle is used to visualise the posterior distribution of the selected terms based upon a Bayesian network model which is constructed for each quarter in order to investigate the causal relationships and interactions between the selected variables within each quarter's network.

5.8.2 Performance Comparison.

Table 5.12 presents the optimum feature attributes selected under the wrapper method for each classifier in each quarter independently, along with their average classification accuracy. Best first search was applied to Bayesian classifiers using the K2 algorithm (Kohavi and John, 1997). A ten-fold cross validation is applied on the whole dataset, where the last column in Table 5.12 represents predictions without feature selection (full feature set). The experimental results interestingly demonstrate that all classifiers perform well under the wrapper method in all quarters. This indicates that the wrapper approach, as a feature selection method, resulted in statistically significant improvements in classification performance over the use of the full feature set of all classifiers. Compared with the other machine learning algorithms, the Bayes net classifiers proved successful and can provide higher prediction accuracy.

Table 5.12: The experimental results of the feature selection and related average classification accuracy of (BN, NB, RandF and SMO) classifiers for all quarters (Qs)

(A) The Performance of four different classifiers on the first quarter (Q1)				(C) The Performance of four different classifiers on the third quarter (Q3)			
Classifiers	Attribute Selected	Classification Accuracy (%)		Classifiers	Attribute Selected	Classification Accuracy (%)	
		Selected Attributes	All Attributes			Selected Attributes	All Attributes
BN	24	69.96	66.05	BN	41	69.35	63.25
NB	28	69.96	65.64	NB	27	68.13	63.49
RandF	34	70.78	62.35	RandF	23	68.38	62.03
SMO	21	67.48	64.2	SMO	15	64.59	64.84

(B) The Performance of four different classifiers on the second quarter (Q2)				(d) The Performance of four different classifiers on the fourth quarter (Q4)			
Classifiers	Attribute Selected	Classification Accuracy (%)		Classifiers	Attribute Selected	Classification Accuracy (%)	
		Selected Attributes	All Attributes			Selected Attributes	All Attributes
BN	22	67.72	61.96	BN	39	72.74	70.16
NB	28	69.45	63.4	NB	38	72.17	69.92
RandF	21	68.3	58.21	RandF	30	70.8	67.26
SMO	20	65.71	63.11	SMO	14	69.03	69.68

Feature Selection and Bayes Net Classifier. Since Bayes net classifiers proved effective in predicting sentiments of StockTwits data, it is worth pointing out at this stage the nature and type of the features selected in each quarter. Table 5.13 presents the wrapper-selected features using the Bayes net classifier of each quarter individually.

Table 5.13: Feature subset selected under Bayes Net classifier for individual quarters

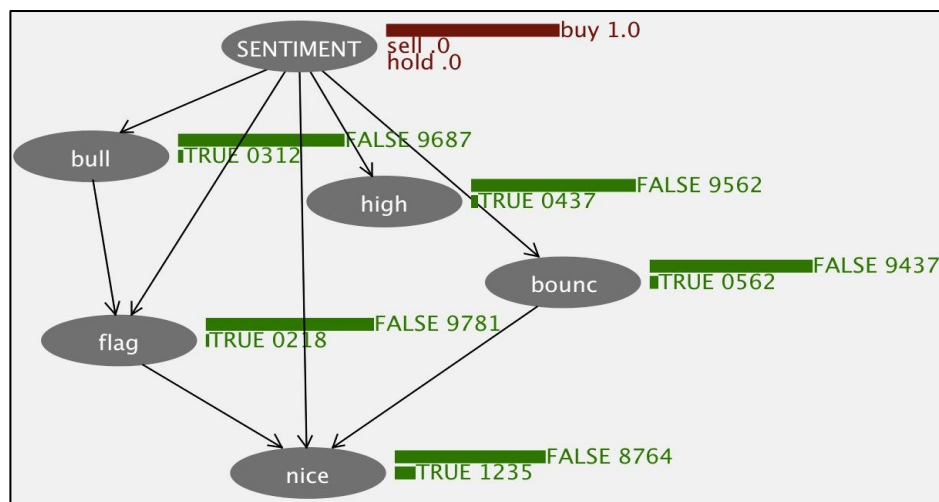
Time	No of Selected Features	Selected Features Subset
Q1	24	aapl , bearish , bottom, bounc, bull, cat , flag, goog, head, high , jnj, lower, nice , sell, set, short , sold , stop, strong, support, weak, xom.
Q2	22	bac, bearish , bottom, bullish, cat , cscoc, current, cvx, high , jnj, jpm, look, low, move, nice , nke, short , sold , time, top, wmt.
Q3	41	bearish , bottom, bought, bounc, break, bull, bullish, buy, call, cat , channel, china, close, continu, cvx, dis, don, high , intc, jnj, jpm, level, lower, move, mrk, nice , nke, pfe, posit, quot, report, sell, short , sold , start, stop, strong, target, unh, xom,
Q4	39	bearish , bought, break, bull, bullish, call, cat , channel, close, cscoc, current, cvx, day, earn, entri, expect, gap, goog, high , hit, ibm, look, lower, mcd, nice , nke, posit, price, report, resist, short , sold , strong, support, trend, utx, via, volum.

Chapter Five: Text Mining Analysis and Findings

As it can be seen from Table 5.13, a number of features appear in almost all quarters (see words in bold) while other features tend to appear in some of the quarters but not in others. An interesting observation from Table 5.13 is that some companies reappeared frequently in some quarters, such as Nike, Inc. “nke” and Chevron Corporation “cvx” and Johnson & Johnson “jnj”, indicating that these companies were highly discussed in the StockTwits forum during that period. This suggests that new information about those discussed companies (e.g. earnings announcements) may be arriving in the market. Claburn, (2009) argues that, as messages are generally posted just before an event occurs, the forum may contain real-time information that is important for making investment decisions.

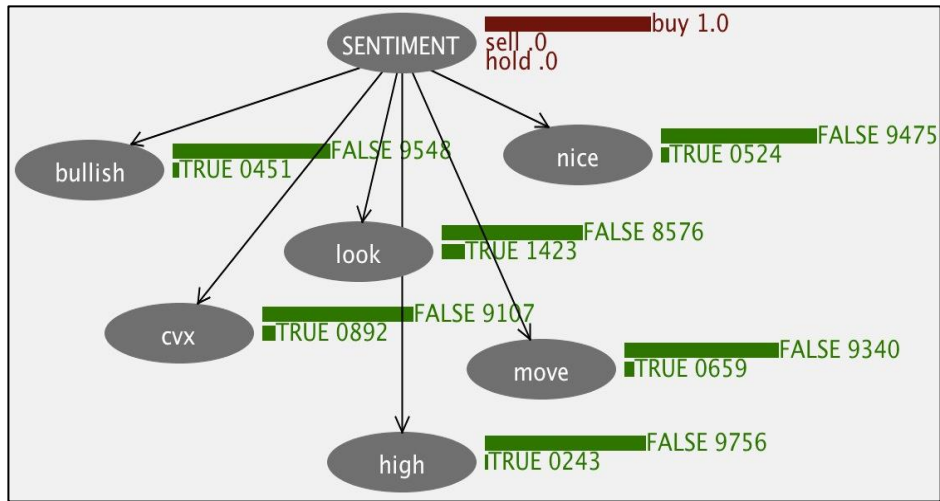
5.8.3 Bayesian Network Model for Sentiment Prediction

Bayesian Networks are built based on the selected features under the wrapper method for four datasets, one for each quarter. Each node in the network represents a term or word that exists in the tweet data whilst the class represents the sentiment. All term nodes in the networks are binary, i.e. having two possible states, which will be denoted by T (True = feature appears in the tweet) and F (False = feature does not appear in the tweet) whilst the class can take on buy, hold or sell states. Figure 5.3 shows extracted versions of Bayesian Networks of each quarter, where the decision/sentiment “buy” is observed, giving a probability value of 1.

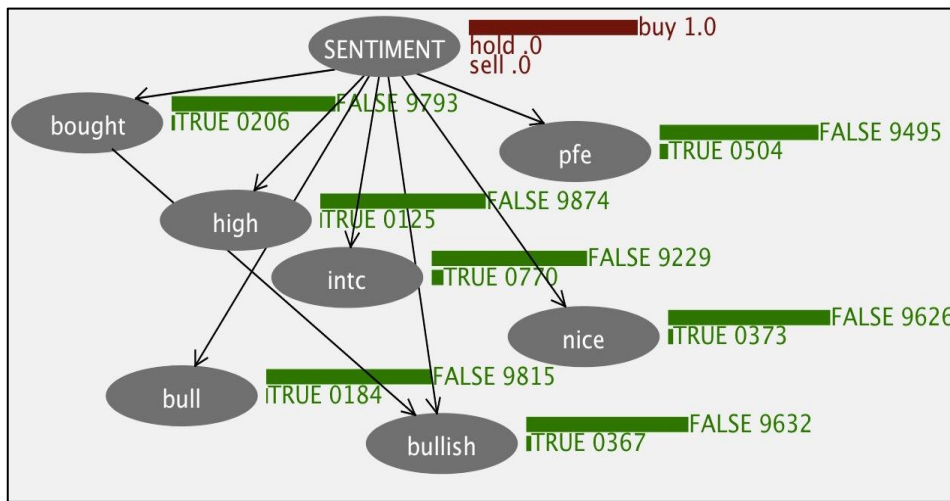


(A)

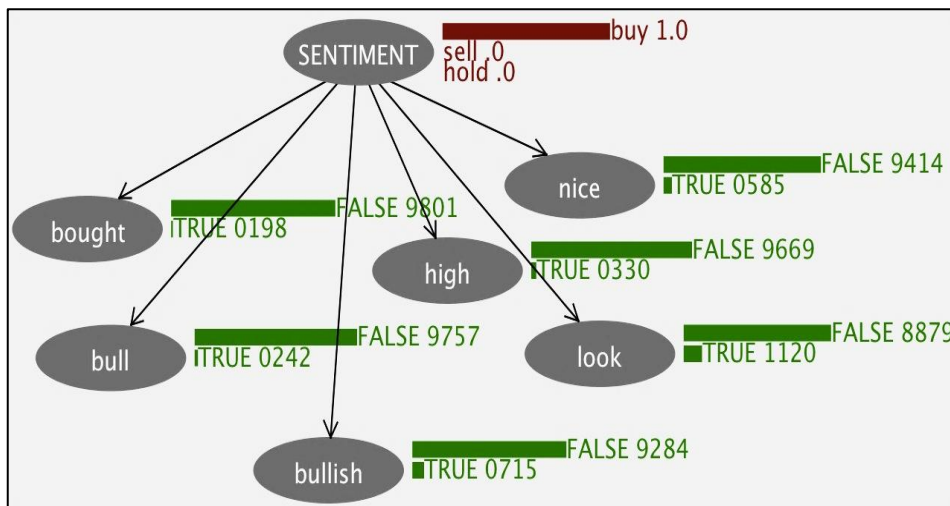
Chapter Five: Text Mining Analysis and Findings



(B)



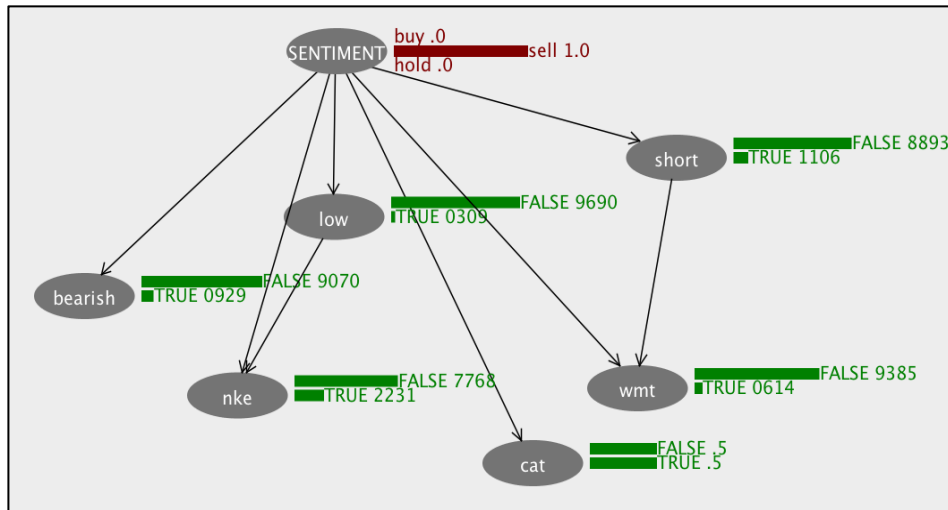
(C)



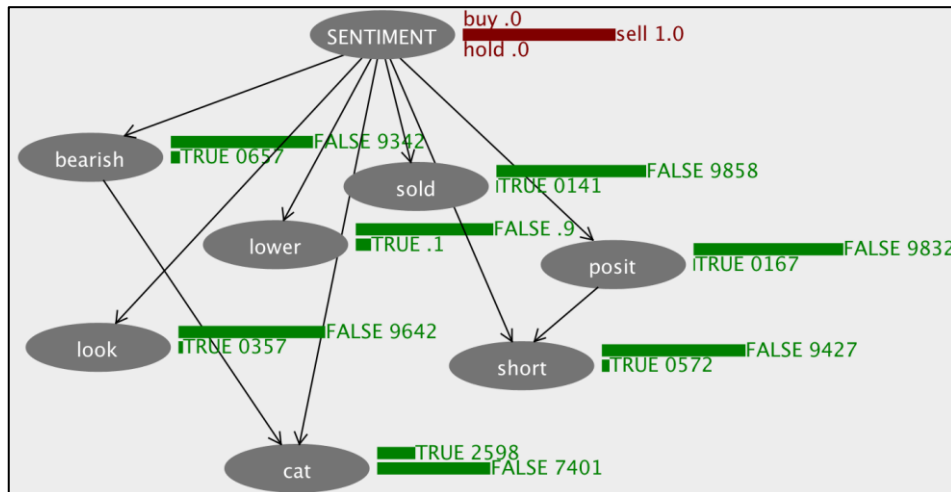
(D)

Figure 5.3: Results of an extracted Bayesian Networks Model of Buy sentiment for (a) Q1, (b) Q2, (c) Q3 and (d) Q4 showing the most dominated words associated with Buy sentiment.

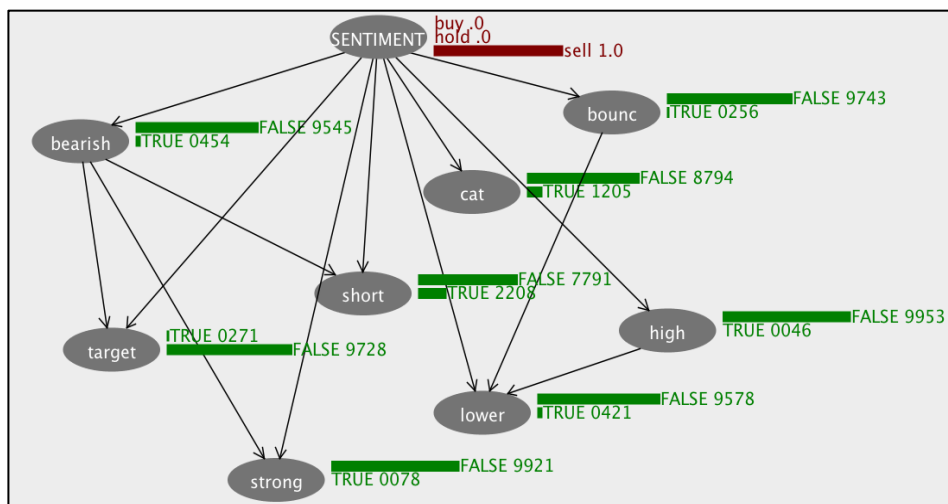
Chapter Five: Text Mining Analysis and Findings



(B)



(C)



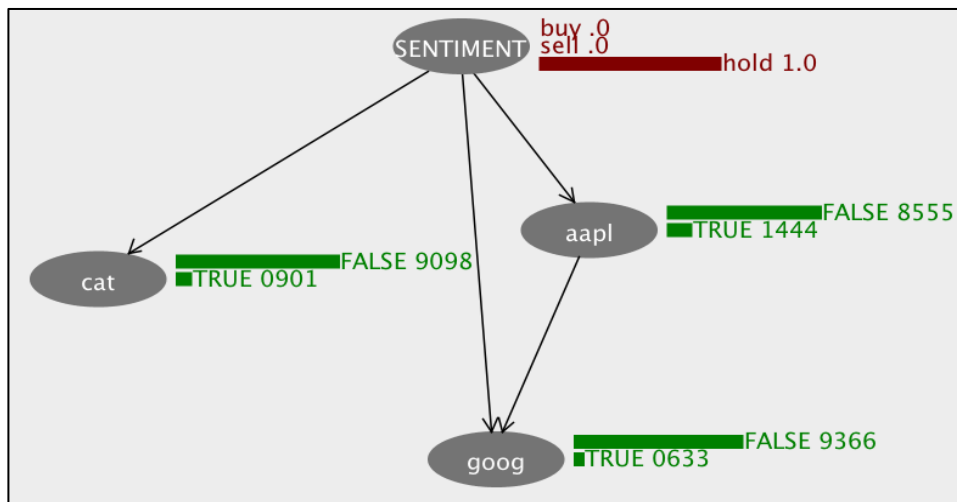
(D)

Figure 5.4: Results of an extracted Bayesian Networks Model of Sell sentiment for (a) Q1, (b) Q2, (c) Q3 and (d) Q4 showing the most dominated words associated with Sell sentiment.

Chapter Five: Text Mining Analysis and Findings

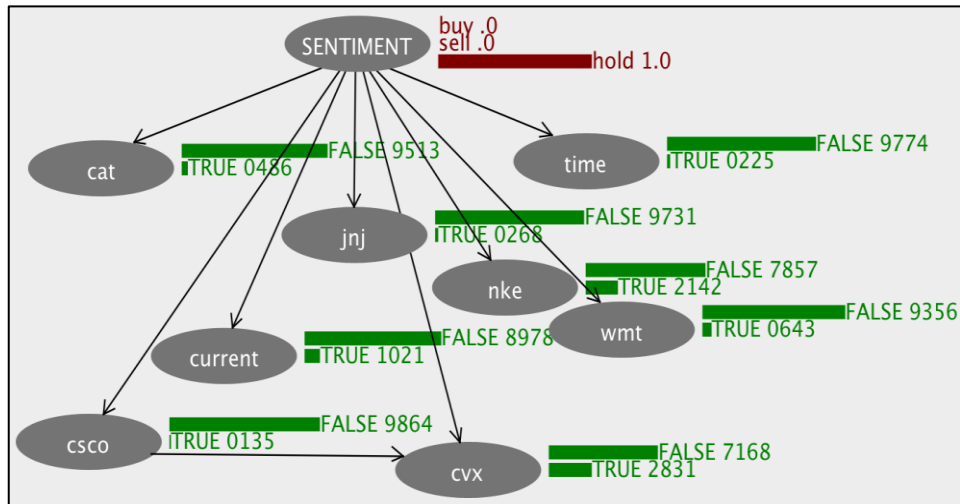
Since the sentiment event has three different states (buy, sell or hold), those words will affect each state differently based on the related weighted probability of their appearance. A simple example can be found in the 1st quarter where “bearish” and “cat” are two child nodes connected with a parent (sentiment). It can be seen that the probability of a sentiment occurring when both features appeared as $P(\text{Sell} | \text{Bearish}, \text{Cat}) = 0.75, 0.056$ and 0.5 for buy, sell and hold sentiments respectively, which means that when both words (bearish and cat) appeared together in a StockTwit message there is an excessive buy sentiment despite their individually prominent appearances in the sell sentiment. Therefore, a sentiment can sometimes be affected inversely depending on whether each word appears independently or in combination.

For the “Hold” sentiment, it can be observed that some words are always likely to appear when the holding sentiment is “on”, indicating either a company’s ticker symbols (e.g. Chevron Corporation “cvx”, Johnson & Johnson ”jnj”, Pfizer, Inc “pfe”) or some neutral words (e.g. report, level). This observation is seen throughout the period, as shown in Figure 5.5 which presents an example of the Bayesian networks model for all quarters (Q1, Q2, Q3 and Q4) when the decision “hold” is perceived, given by a probability value of 1.

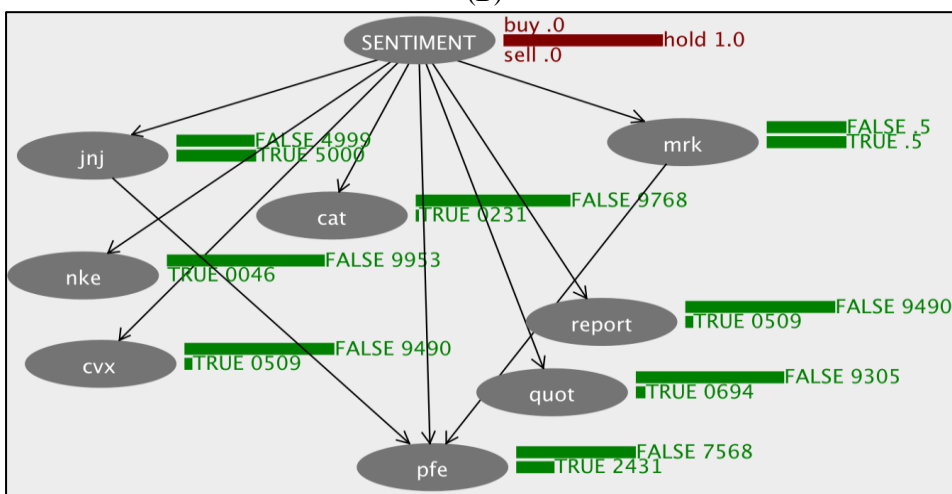


(A)

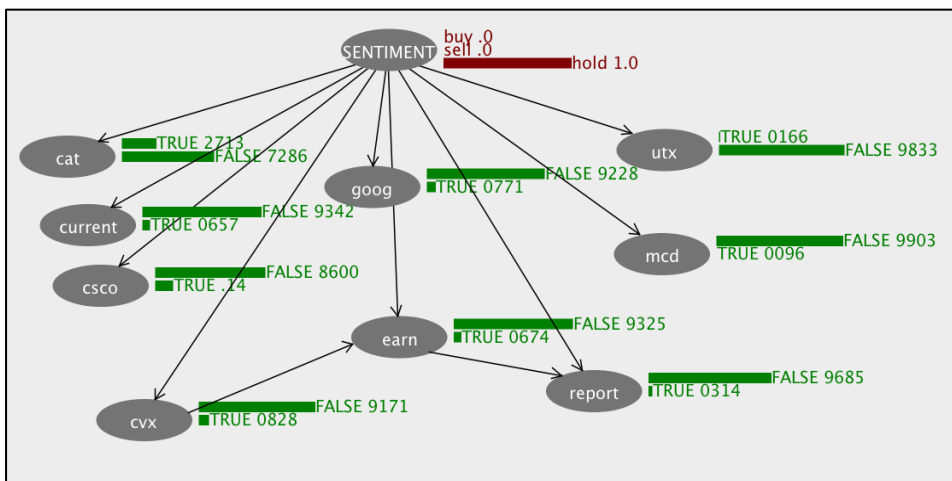
Chapter Five: Text Mining Analysis and Findings



(B)



(C)



(D)

Figure 5.5: Results of an extracted Bayesian Networks Model of Hold sentiment for (a) Q1, (b) Q2, (c) Q3 and (d) Q4 showing the most dominated words associated with Hold sentiment.

Chapter Five: Text Mining Analysis and Findings

The change in conditional probability distributions of the most prominent words associated with the buy, sell and hold sentiment over time are shown in Figure 5.6

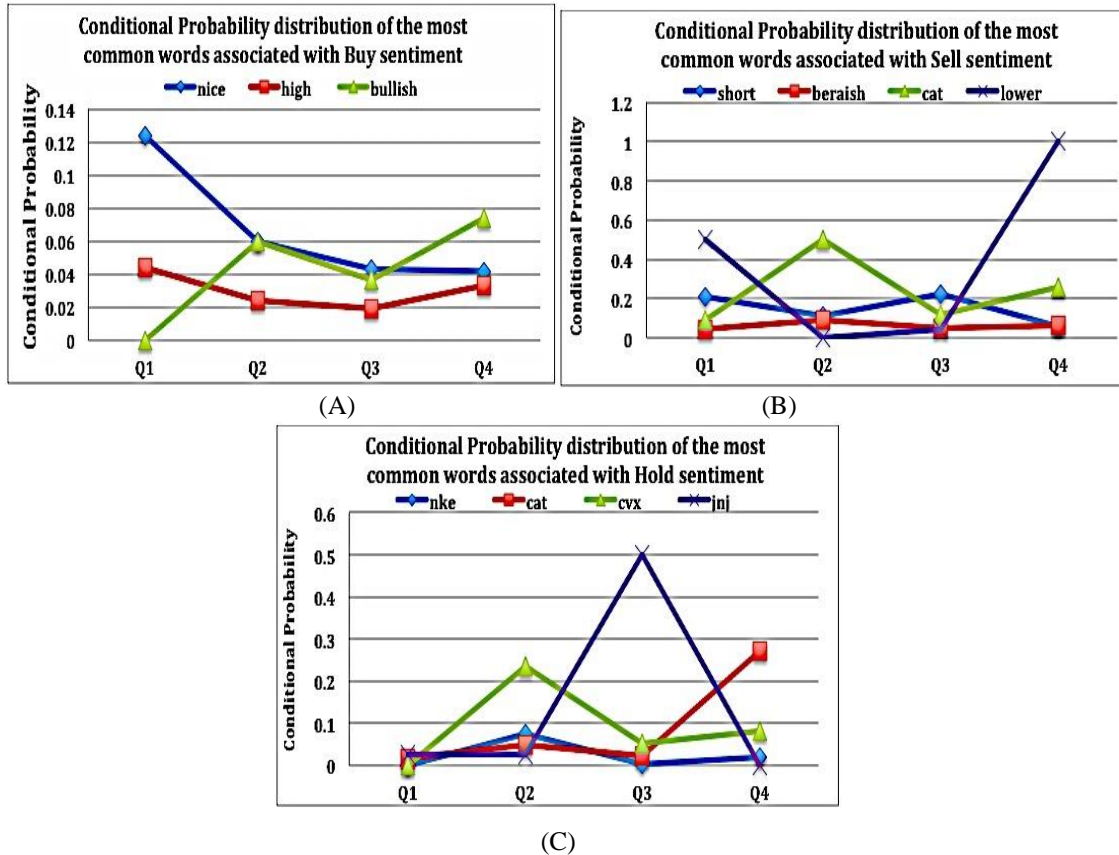


Figure 5.6: The conditional probability distribution of the most common words related with the (A) buy, (B) sell and (C) hold sentiment

5.8.4 Textual Visualization of features Selection Using Wordle

Wordle is a text analysis tool used to highlight the words that most commonly occur throughout StockTwits text (Wordle, <http://www.wordle.net/creat>). It creates an image that randomises the words where the size of the words is determined according to the frequency with which they occur, highlighting their importance. In our case, the probability values of all features, which are obtained from a Bayesian network, are used to determine the prominence of those features. Figure 5.7 shows the visualised image of the selected features that are more associated with particular sentiments in all quarters.

Chapter Five: Text Mining Analysis and Findings

“dis” and “intc” are companies clearly visible in sell sentiments indicating a high bearishness where investors might tend to sell short their stocks of those companies. The textual visualisation windows of the “hold” sentiment interestingly show that greater prominence is given to words that represent the company ticker symbols as well as some other words (e.g. “report”, “qout”, “don” and continue”), throughout the quarters. For example, the most dominant words associated with the “hold” sentiment are “aapl”, “goog” and “xom”, “csc0”, “cvx” and “wmt”, suggesting that these corporations are mostly being held during the 1st and 2nd quarters. However some corporations reappear and demonstrate a holding position, especially the largest corporations such as “csc0” and “goog” which always seem to be associated with hold messages.

5.8.5 Discussion

The experiments proved the predictive ability of Bayes Net classifiers in predicting StockTwit sentiment while Bayesian Networks models and textual visualisations together provided a very useful graphical representation of the feature selected under wrapper method of feature selection. In general, a look at the most prominent words per sentiment class indicate that the Bayesian Network model and textual visualisation using Wordle derived a plausible dictionary from the training set. Obviously, some features occur frequently in all sentiment classes (e.g. look). The positive emotions (e.g. nice and strong) are much more likely seen in the buy sentiment, while the sell sentiment contains much more negative emotions (e.g. “stop”, “low” and “close”). Buy sentiment reflects the linguistic bullishness and more likely contain “bullish” word along with other technical words (e.g. “move” and “high”) or trading words (e.g. “buy”, “bought”, “bull” and “call”). On the other hand, the sell sentiment reflects bearishness and often combines the “bearish” word with technical words (e.g. “support”, “lower”) or trading words (e.g. “sell”, “sold” and “short”). Hold sentiment more likely contains neutral words (e.g. “report”, “quote” and “time”) or company names (e.g. the company ticker symbol; “cvx”, “csc0” and “jnj”). An equal balance of negative and positive emotions are likely to be found in the hold sentiment. Based on what the findings of this research on the ability of Bayesian Network in predicting investor’s sentiment in the stock market, this may yield

Chapter Five: Text Mining Analysis and Findings

promising insight into the potential provision of an investment support mechanism for analysts, investors and their peers. Practically, this could be used to determine the accurate time when stocks are to be held, when to be added (buy) and when to be removed from a portfolio that yields maximum return on investment for the investor.

5.9 Application (2): Application of Filter Approach: Quantifying StockTwits Semantic Terms' Trading Behaviour in Financial Markets: An Effective Application of Decision Tree Algorithms

The provision of an accurate and timely trading support mechanism is the key to success for traders seeking to make profitable decisions in capital markets. This study presents a novel approach for developing a new decision support system based on tweet semantic terms extracted from the decision tree model (Quinlan, 1993), which can then be implemented as a trading strategy. It constitutes three different portfolios (sell, buy and hold). The decision tree proved successful in searching for rules hidden in large amounts of data. The visibility of the relationships between nodes, branches and leaves in the tree makes it the most suitable approach for feature selection and prediction of investment trading decisions in capital markets. It has also proved efficient for time-series analysis. In addition, decision tree techniques have already been shown to be interpretable, efficient, problem-independent and capable of dealing with large-scale applications. The decision tree model provides a visualised insight into the StockTwits data by highlighting the individual relationships with respect to the class as well as the combined associations of features with respect to the decision class. One would expect the decision effect of individual terms (features) appearing in a tweet posting to be different from that produced had it appeared in combination with other terms. The ability of the decision tree model to explore the related interactions between the selected terms and their ability to predict trading decisions makes it a better and more suitable model for this research. This research aims to predict an intelligent trading support mechanism to screen out the most significant and profitable trading terms or combination of terms from StockTwits data that may help investors to make correct and accurate (selling, buying or holding) decisions in capital markets. The research attempts to investigate whether the terms or combination of terms of trading decision rules extracted from the decision tree algorithm will act as a trading decision guide for investors that may lead to profitable

Chapter Five: Text Mining Analysis and Findings

investment decisions while examining the predictive ability of each term or combination of terms in anticipating subsequent movements in the stock market.

5.9.1 System Pipeline

The System pipeline diagram illustrated in Figure 5.8 outlines the methods employed for this application. There are essentially five phases; Data acquisitions, Text Processing Model, Feature Selection, Performance Evaluation and Portfolio Construction and Investment Hypothesis. These phases are represented in dashed boxes identified with the relative name marked in red. Each component framework consists of different procedures that are vital in performing the whole function of the relative phase.

The *Data Acquisitions and Pre Processing* phase is the first component that appears at the top left of the figure, which is accountable for data description and procurement from various sources as well as pre-processing and filtering procedures to avoid irrelevancy of the data being collected. At this stage and after the text customization has been performed¹⁸, the manual operation of sample tweet messages is performed to manually classify tweets into three distinct class namely; sell, buy or hold using the Harvard IV dictionary which then is used as a training set in the text processing model and feature construction stages¹⁹. The second component, *Feature Selection*, represents the implementation of filter approaches of feature selection (based on Information Gain criteria (IG)) to extract the most relevant features from the datasets to build a features construction model. The construction model of relevant features (reduced features) is then used as input variables to the third component *Text Processing Model* where the Decision Tree algorithm C4.5 is employed to process the text and detect relative sentiments. The trading decisions rules: sell, buy or holds of each term or combination of terms are extracted from the decision tree classifier. The proposed system treats each term (or combination of terms) as a trading strategy called Tweet Term Trading (TTT) Strategy and calculates the cumulative return from such strategies accordingly. These trading strategies are then evaluated by comparing its performance to a benchmark trading strategy (e.g. Random Strategies, Buy and

¹⁸ There are a number of customisations that have to be performed at this stage to maximize the classification performance. This includes preprocessing steps like: stop-word removal, stemming procedures, removal of unnecessary words, tweets have to be in lowercase and text reformatting (e.g. whitespace removal)

¹⁹ More details about this phase are provided in the framework design in previous chapter (chapter four).

Chapter Five: Text Mining Analysis and Findings

Hold Strategy and Dow Jones Strategy), which is the task handled in the fourth component of the design *Performance Evaluation*.

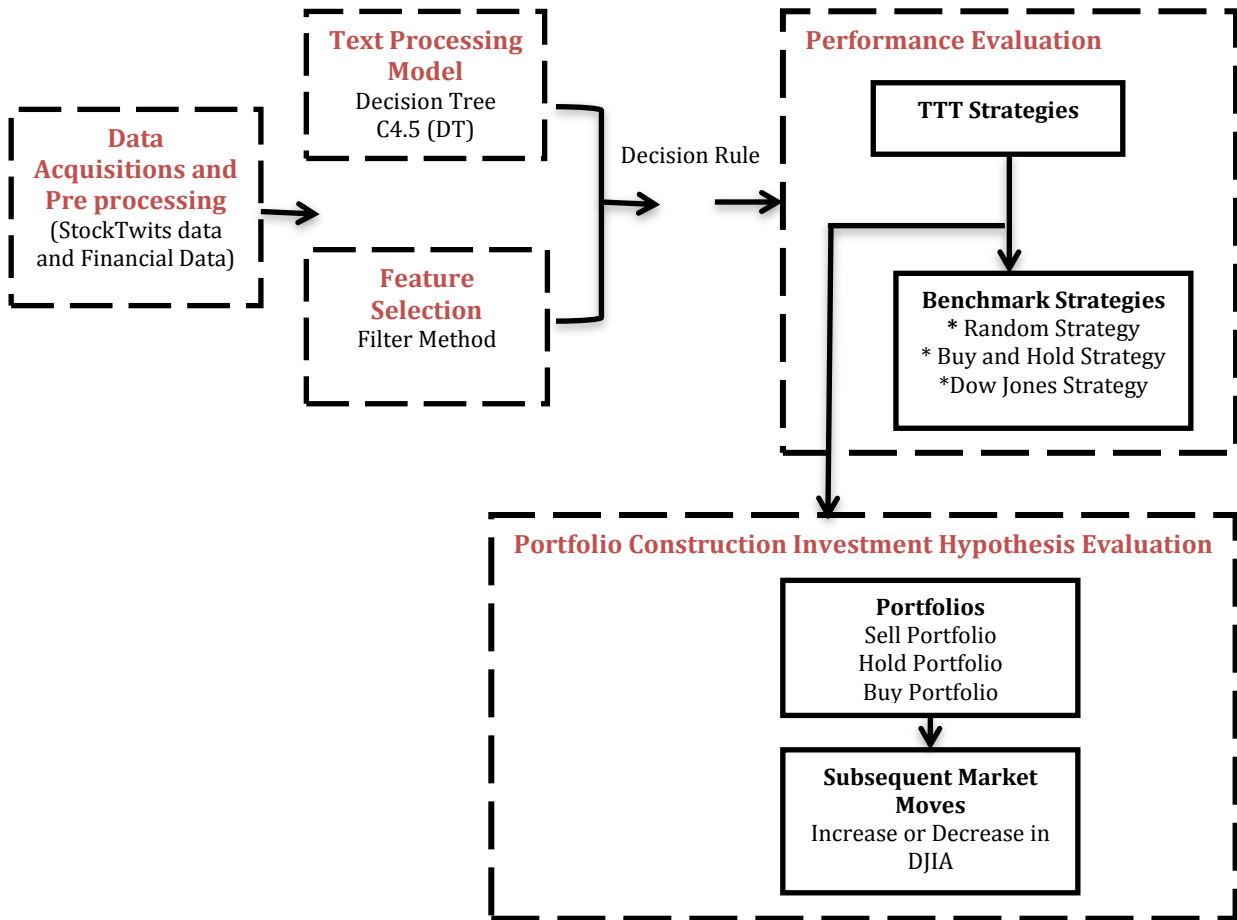


Figure 5.8: System Pipeline

In the final component *Portfolio Construction and Investment Hypothesis Evaluation*, investment portfolios for each decision class (sell, buy and hold) are constructed where each portfolio consisting of all possible terms and/or combination of terms belong to that class. Moreover, the investment-trading hypothesis (short and long position) adopted is empirically tested to calculate the cumulative return for each trading strategy.

5.9.2 Trading Strategies Design

Widespread evidence has been growing that stock prices overreact or underreact to information which suggests that a profitable trading strategy that selects stocks based on their past returns will probably exist. The concept of this research is

Chapter Five: Text Mining Analysis and Findings

built upon the previous research study of Tetlock et al. (2008), who found that a trading strategy based on negative words in firm specific news articles could earn abnormal annualised returns. To more thoroughly test the ability to earn abnormal profits based on specific terms in StockTwits messages, a trading strategy was designed as introduced in Preis et al. (2013), for some specific terms or set of terms that are believed to have an effect on the selling, buying or holding decisions in capital markets (as suggested by the feature selection method and the decision tree algorithm discussed earlier in this chapter). Unlike the study of Tetlock et al. (2008), who used a simple quantitative measure of language to predict firms' accounting earnings and stock returns based on negative words alone, this study considers a collective use of the tweet language whereby positive, neutral and negative words are all considered in predicting tweet term trading strategies.

To investigate whether the occurrence of a specific term or combinations of terms have the power to predict a trader's decision in a capital market, closing prices $p(t)$ of the Dow Jones Industrial Average (DJIA) were analysed on a daily basis over a one year period. In this strategy, StockTwits data are used to obtain a volume frequency $n(t)$ of a term in day t . Then, a daily time series is created for the terms and/or the combination of terms based on the daily volume frequency of terms that appears in the tweet messages over the studied sample period. In the non-trading days, the volume frequency of a given term/combination of terms will be combined together with the volume frequency of the next immediate trading day. Note that there might be a silent period either because there were no messages posted or the terms might not have appeared in that particular tweet posting. In line with the study of Antweiler and Frank (2004b) on the Internet message board, all silent periods are placed with a value of zero. To minimise the effect of the silent period, focus is only on the terms with high volume frequency of appearance by ensuring that the minimum value of the term frequency considered is no less than 100, which represents a minimum volume frequency of the terms considered in this study. To compare the changes in term volume frequency to subsequent market moves, a trading strategy for each of the 122 terms is implemented. The following section will explain the design of the proposed trading strategy followed in this research. To quantify changes in the appearance of a term in a tweet message, the relative change in volume frequency is used:

Chapter Five: Text Mining Analysis and Findings

$$\Delta n(t, \Delta t) = n(t) - N(t-1, \Delta t) \quad (5.1)$$

where $n(t)$ = the volume frequency of a term appeared in a given day and $N(t-1, \Delta t) = (n(t-1) + n(t-2) + \dots + n(t-\Delta t)) / \Delta t$ is the average number of term frequency of the previous 5 days. This method is called a simple moving average (MA) method where it is used to roll out the effect of the term appearance over the previous five days average. The term frequency over five realizations of its frequency value is averaged assuming that the effect of that term will last at least five trading days.

The proposed trading strategy presented in this thesis is called tweet term trading (TTT) strategy. It simply evaluates the profitability of a tweet term strategy and is substantially effective for investors as it provides guidance in helping make a correct, accurate and profitable decision concerning a particular security in a capital market. As it is well known, a trading strategy makes profits only if it could provide some predictability of future changes in stock prices, given the great variability of the data in the stock market. Therefore, the investment strategy is evaluated by hypothetically implementing it as follows:

$$Stat(t) = \begin{cases} \text{Short position,} & \text{If } \Delta n(t-1, \Delta t) > 0 \\ \text{Long position,} & \text{If } \Delta n(t-1, \Delta t) < 0 \end{cases} \quad (5.2a)$$

$$Sig(t) = \begin{cases} \text{Short position then, sell } p(t) \text{ and buy } p(t+1) \text{ and } Rtn = Ln p(t) - Ln p(t+1) \\ \text{Long position then sell } p(t+1) \text{ and buy } p(t) \text{ and } Rtn = Ln p(t+1) - Ln p(t) \end{cases} \quad (5.2b)$$

Stat (t) denotes the current trading position of investors, while sig(t) indicates the trading instruction produced in this strategy design. According to this strategy, investors take a short position in the market following an increase in term volumes frequency ($\Delta n(t-1, \Delta t) > 0$) by selling the DJIA at the closing price $p(t)$ on the first trading day and buying back the DJIA at price $p(t+1)$ at the end of the following day. If instead a long position has been taken following a decrease in term volume frequency ($\Delta n(t-1, \Delta t) < 0$) then investors buy the DJIA at the closing price $p(t)$ on the first trading day and sell the DJIA at price $p(t+1)$ at the end of the next trading day. A cumulative return for each trading strategy therefore needs to be calculated. If investors take a 'short position', then the cumulative return R is $Ln p(t) - Ln p(t +$

Chapter Five: Text Mining Analysis and Findings

1) whereas, if he/she takes a 'long position', then the cumulative return R then changes by $\ln p(t+1) - \ln p(t)$. Following this strategy, it is assumed that buying and selling activities will have a symmetric impact on the cumulative return R of a strategy's portfolio. As usual in this type of analysis, transaction costs are usually ignored (Zhang and Skiena (2010)). However, one cannot rule out the impact of such transaction costs on impacting profit in the real world implementation. Therefore, this study follows Hu et al. (2015) by considering the transaction cost to evaluate the performance of the (TTT) strategies proposed here. Clarkson et al. (2006) argues that the level of transaction costs for online brokers are in the range of 0.15%-0.2%.²⁰

5.9.3 Benchmark Trading Strategies

To assess the profitability of the tweet term trading strategies created in the previous section; the performance of these strategies has to be evaluated against benchmark trading strategies. The purpose of this research is to find out whether the trading strategies based on the semantic terms in StockTwits forums could earn abnormal profits, while it is not being emphasised here that these strategies are the optimal and the best strategies for investors. In the present study, three benchmark trading strategies are considered as described in the following subsections.

1. *Random (RND) Strategy*

Random investment strategy is the simplest strategy where at time t the correspondent trader makes his/her prediction on trading completely at random. An investor following such a strategy makes decisions each day to sell or buy the market index in an uncorrelated, random manner. In any given day, there is an equal chance (probability = 50%) that the index will be bought or sold and this decision is independent and unaffected by decisions in the previous day. Statistically speaking, random strategy is a normal distribution strategy with mean value of $\langle R \rangle_{Random\ Strategy} = 0$. In trading analysis, the means of any trading strategies developed are tested against the mean of the distribution curve that a random trading strategy would produce, which in statistics is assumed to be zero under the null hypothesis of no excess returns. (Vanstone and Hahn, 2010). As with any standard

²⁰Tetlock et al. (2008) even use only 10 bps to assume reasonable transaction costs.

Chapter Five: Text Mining Analysis and Findings

normal random variable, the standard deviation of this strategy is derived from simulations of 1,000 independent realisations of uncorrelated random strategy as shown in Figure 5.9.

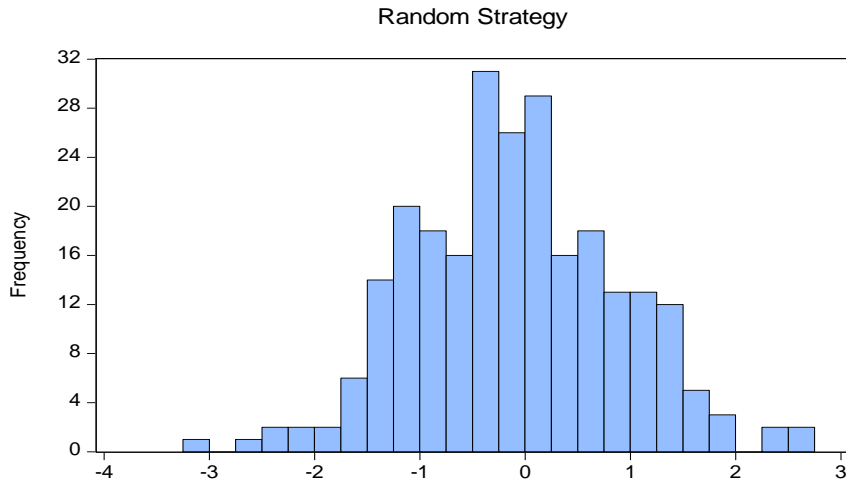


Figure 5.9: The standard deviation of 1,000 simulations of average returns using purely random investment strategy.

2. Buy and Hold Strategy

Buy and Hold strategy is defined as a passive investment strategy in which investors take a passive role in the market with no active buying and selling of stocks from the time the portfolio is created until the end of the holding period (end of investment horizon). The 'buy and hold' strategy is implemented by buying the index at the beginning of the period 3rd April 2012 and selling it at the end of the holding period of investment at 5th April 2013. This strategy yields 10.347% profit, which is equal to the overall increase in value of DJIA over the investment period of one year from 3rd April 2012 until 5th April 2013. The return obtained from this strategy is 0.0985 standard deviations of cumulative returns of uncorrelated random investment strategies.

3. Dow Jones (DJ) Strategy

This strategy is based on changes in DJIA prices $p(t)$ instead of changes in the term related frequency data as the basis of buy and sell decisions. Implementing this strategy resulted in a loss of 6.177%, or when determined by the mean value of random strategy, results in a negative return of -0.0245.

5.9.4 Empirical Test and Analysis

A) Filter Approach For extracting the filter subset, a ranker search method was used (Mark et al., 2009) in conjunction with the information gain criteria where the worth of an attribute is evaluated by measuring its information gain (IG) score with respect to the class. Referring to Table 5.4 in Section 5.4, 45 terms are retained from performing the filter approach using information gain criteria. The terms listed in the table are ranked according to their relevancies where the terms at the beginning of the list (indicated by the serial number) are most relevant, as the relevancy decreases as one goes down the list. The IG value is reported next to each term. For example, the term ‘short’ appears to be the most significant term among all listed terms with IG value of 0.0706 while ‘run’ is the least important term with IG value of 0.0034.

B) Decision Tree Model Quinlan’s C4.5 (DT) algorithm (Quinlan, 1993), is used to classify the tweet messages based on the reduced model of the features selected under a filter approach using the IG criterion. Performing feature selection using decision trees reveals that 45 attributes, indicated by the nodes in the tree model, are regarded as the most relevant features that can make a better prediction of the three decision classes (buy, sell, hold). All of the selected features were deemed relevant in predicting the sentiment class, whether these feature nodes connected directly to the decision class or were connected through leaves with other decision nodes in the tree to the sentiment class. One of the main advantages of the decision tree model is that it naturally explores interactions between terms via the visualised connections between different nodes connected through leaves in the decision tree. To provide more understanding of the connected relationships between terms in the tree model, an extracted version of the visualised tree is provided while explaining the nature and type of these connected relationships for classifying StockTwits sentiment class. Figure 5.10 shows the visualised output of some of the selected features of the decision tree near to the root node. Other aspects of the extracted versions of the tree model are shown in Figure a in Appendix VI.

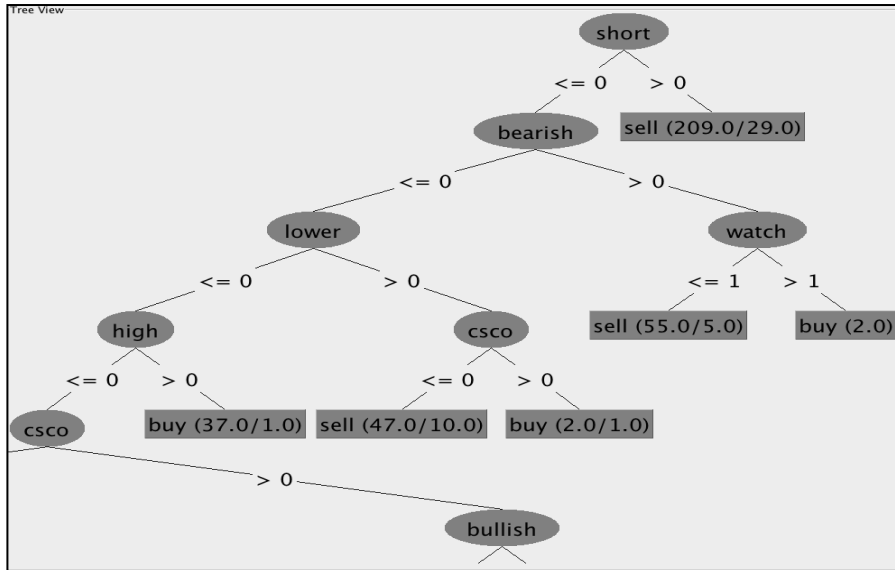


Figure 5.10: An extracted version of the visualised decision tree model

From the extracted visualised tree model graphed in Figure 5.10, it can be seen that the decision node (sell) is connected, through leaves, to the words (short, bearish and lower). This indicates that these words are the most relevant words that best classify “sell” messages. Each term is indicated by a node in the tree and connected through leaves to one of the three decision classes (sell, buy or hold). In some cases, a set of terms might be connected together to one decision class, where this indicates that the combined appearance of these connected terms may have a different effect on the trading decisions than when the term appears alone. For example, when the decision node “bearish” appears in a tweet message, it indicates a sell decision as it is connected through leaves to the decision class sell. However, when the term “bearish” is connected together with the decision node “watch” through leaves, it indicates a buying decision despite its individual independent appearance as a sell decision. Therefore, trading decisions can sometimes be affected inversely depending on whether each term appears independently or in combination. Due to the large size of the decision tree generated for StockTwits data in this research, another exemplary screen of a visualised decision tree will be provided in Figure b in Appendix VI.

A set of decision rules can be generated from the DT model by following the decision tree from top to bottom. These decision rules are based on the idea that the appearance of a term or a set of terms in tweet postings might inform investors about whether to buy, sell or hold a stock in a capital market. Therefore, it is worth pointing

Chapter Five: Text Mining Analysis and Findings

out at this stage the nature of the decision rules that can be extracted from the DT model. Table 5.14 shows the trading decision rules corresponding to each term that can be extracted from the decision tree model. Note, that the individual appearance of terms indicating the company ticker symbols is excluded as shown in bold in Table 5.4 That is because including the single appearance of such terms might bias the volume frequency, which may result in misleading the strategy performance of such terms. However, the combined appearance of those terms is still considered, as they might be more informative when appearing together with other terms in the tweet postings.

Table 5.14: The decision rules for individual occurrence of the term in the StockTwits postings

<i>Decision rule:</i> If the term	<i>bearish</i> <i>botoom</i> <i>bounc</i> <i>flag</i> <i>lower</i> <i>sell</i> <i>short</i> <i>stop</i> <i>support</i> <i>volume</i>	appears in a tweet message then the decision would be Sell
<i>Decision rule:</i> If the term	<i>current</i> <i>entri</i> <i>market</i> <i>post</i> <i>report</i> <i>set</i> <i>week</i>	appears in a tweet message then the decision would be Hold
<i>Decision rule:</i> If the term	<i>bought</i> <i>break</i> <i>breakout</i> <i>bull</i> <i>bullish</i> <i>buy</i> <i>head</i> <i>high</i> <i>look</i> <i>move</i> <i>nice</i> <i>run</i>	appears in a tweet message then the decision would be Buy

Table 5.14 shows that there are some specific terms associated with the decision classes; sell, hold and buy where their appearance in a StockTwit message gives indications to financial market practitioners as to whether to sell, buy or hold the discussed stocks. For example, if terms like “bought”, “bullish”, “move” and

Chapter Five: Text Mining Analysis and Findings

“nice” appear in a tweet posting discussing a particular stock of DJIA, that provides a buying signal to investors to buy that particular stock. While the appearance of terms like “bearish”, “bottom”, “lower” and “short” indicates a sell signal to investors and most probably recommends investors to take a sell decision concerning that particular stock. The appearance of terms such as “report”, “market”, “week” and “set” seems to inform investors to hold the discussed stocks. Looking closely at the nature of the terms associated with each decision class this study finds that StockTwits postings provide reasonable reflections of the linguistic bullishness of the three classes (buy, sell and hold). This research finds that positive emotional terms are more likely associated with the decision ‘buy’, which by nature reflects investors’ optimism towards particular traded stocks in financial markets. On the other hand, negative emotional terms are likely to be associated with the decision ‘sell’ indicating investors’ pessimism about that particular stock. Neutral terms are more likely to be found in a tweet message discussing a particular stock if a holding decision is to be made by investors.

Having discussed the decision rules associated with the individual occurrences of some terms in the StockTwits postings, it is important therefore to shed light onto the impact of the combined appearance of those terms with other terms in tweet postings. Table 5.15 shows the decision rules obtained from the DT model where it is the set of terms or combination of terms that constitute the decision rules rather than individual terms.

Table 5.15 provides the trading decision rules extracted from the DT model, where a set of rules based on the combined appearance of the terms in tweet postings are listed under the decision class where they belong²¹. What is interesting in Table 5.15 is that the companies’ ticker symbols such as “csc”, “jpm”, “mrk” and many others, when combined with other terms, contain valuable information regarding decisions to be taken by investors not just merely a ticker symbol of the relative company. The most surprising aspect of the decision rules presented in the Table 5.15 is that the trading decision rules differ completely depending on whether the term independently appeared in a tweet message (see Table 5.14) or in combination with other terms. For example, while the appearance of the term “lower” in Table 5.14,

²¹ Note that to maintain unbiased results Table 5.16 only reports the term and combination of terms that have a minimum total volume frequency of 100 over the period studied where the terms/combination of terms of less than 100 value frequency will be withdrawn from the analysis.

Chapter Five: Text Mining Analysis and Findings

indicates a purely sell decision, this term when combined with the company ticker symbol “csc0” markedly indicated a buying position to be taken by investors (see last column of Table 5.15). Another example that demonstrates this finding is when considering the term “look” where its individual appearance indicated a buy signal to market participants, this term when mutually combined with other terms (i.e. “look + intc”, “look + hold”, “look + close” and “look + daily”) excessively signifies a sell signal to investors.

Table 5.15: The decision rules for combinations of terms appeared in the StockTwits postings

<i>Decision rule:</i> If the term “...”, the term “... “ and /or the term “...” appeared in a tweet message then the decision would be		
Sell	Hold	Buy
unh + gap	appl + jpm + amzn	lower + csc0
mrk + amp	csc0 + amzn + goog	bullish + csc0
mrk + break	appl + jpm	csc0 + amzn
report + intc	appl + stock	csc0 +trend
report + wmt	appl +wmt	csc0 + break
break + mmm	appl + market	cvx + move
break + nke	appl + trade	cvx + entry
xom + bottom	appl + msft	cvx + xom
amp + head	goog + sell	cat + mmm
look + intc	sell + pfe	cat + break
stock + jpm	watch + jpm	cat + run
week + nke	watch + nke	cat + call
wmt + qout	watch + wmt	unh + hold
break + bounc	mcd + trade	unh + day
break + stock	stock + amzn	unh + look
break + support	jnj + wmt	appl + nke
break + weak	wmt + friday	report + jpm
bullish + market	chart + post	appl + ibm
chart + flag	market + time	spy + msft
chart + price	watch + follow	axp + ibm
day + expect	watch + list	axp + look
day + news	watch + news	amp + news
head + move		amp + stock
hold + gap		amp + daily
hold + look		amp + sold
look + close		amp + trade
look + daily		amp + dis
stock + current		stock + amzn + wmt
strong + support		jnj+ chart
support + break		bought + sell
week + daily		break + look
week + time		hold + bounc
yesterday + bought		hold + play
		report + low
		stop + current

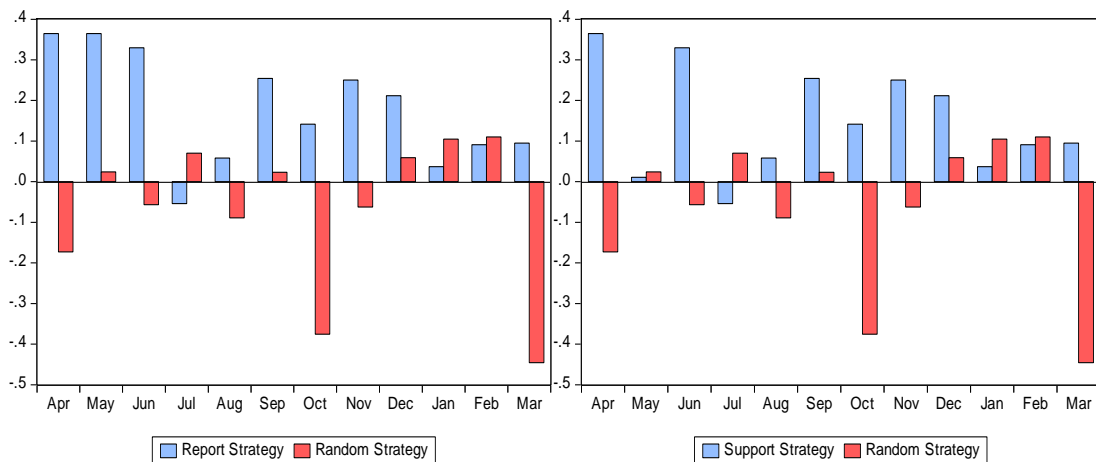
5.9.5 Performance Evaluation

Implementing the trading strategy for all of the 122 trading decisions reveals that the majority of the terms (95 terms trading strategies) outperform the random

Chapter Five: Text Mining Analysis and Findings

strategies indicated by positive returns. However, the remaining 27 terms show negative returns indicating that these strategies fail to perform better than random chance. Constructing the investment strategy defined in equations (5.2a) and (5.2b) for each time series of all the terms/combination of terms presented in Table 5.16, information is provided not only about the cumulative average return but also about the number of the buy/sell signals per TTT strategy. Table 5.16 reports the number of trades per strategy along with its corresponding cumulative returns. As it can be seen from Table 5.16, the first column reports the list of the tweet term, while the average returns and the number of trades of the corresponding term are shown in second and third column respectively. The column reporting the number of trades indicates the total number of the buy and sell signals conducted for each term when implementing the investment strategy defined in equations (5.2a) and (5.2b). The tweet term and/or combination of terms are listed in accordance of performance based on the cumulative average returns.

Evaluating the overall trading strategies reveals that, the term “report” appears to be the best performing term in our analysis followed by the term “support”. Figure 5.11 shows the monthly average cumulative performance of the top four trading strategies: “report”, “ support” “report+intc” and “support+ break”. The blue bars in the graphs depict the cumulative return of our trading strategies where the spikes of these blue bars are more likely pronounced at the top half of the figure indicating positive returns. The red bars on the other hand indicate the standard deviation of the cumulative return from a random strategy (in which buying and selling is done in an uncorrelated random manner) where more spikes of these red bars are pronounced at the bottom half of the graph indicating negative returns in general.



Chapter Five: Text Mining Analysis and Findings

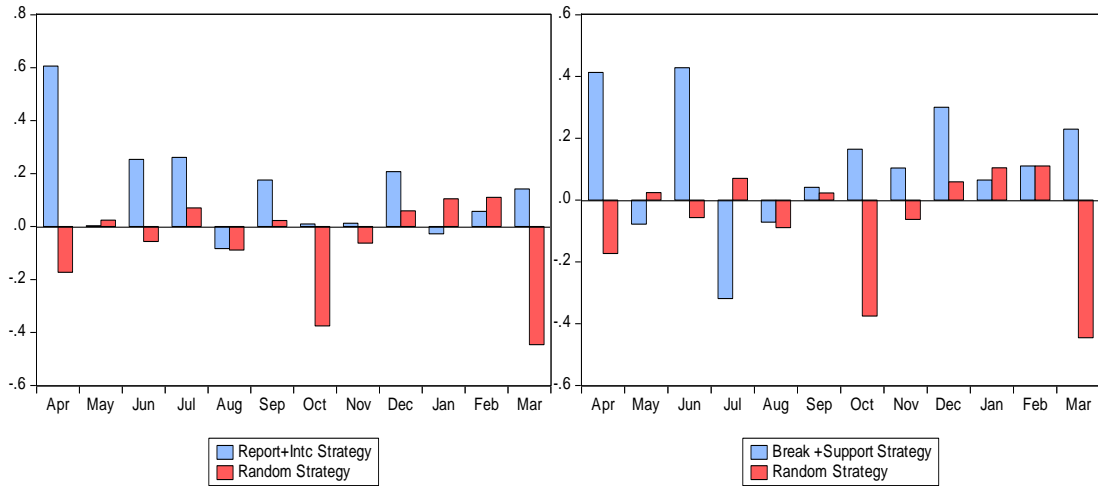


Figure 5.11: A comparison of the monthly average cumulative performances of the trading strategy of the tweet terms “report”, “support”, “report+intc” and “break+support” with the random investment strategy.

From Figure 5.11, it can be seen that the trading strategy of the best four performed terms is performing better than random strategy meaning that there are significant higher positive returns than the random investment strategies in all graphs. As it can be seen from the four charts above that more spikes of blue bars are found in the upper area of positive returns in contrast with random strategy where the red bars spikes more in lower negative return area of the graphs.

The full ranked list of the 122 investigated tweet terms by their trading performance indicated by the cumulative average returns of each strategy. Figure 5.12 depicts the cumulative return of the 122 TTT investment strategies based on their performance. Figure 5.12 shows that the vast majority of the TTT strategies are profitable as these strategies resulted in cumulative average returns greater than the random strategy $\langle R \rangle_{Random Strategy} = 0$. The top half of the figure denoted by the red bars indicates the strategies with positive returns, while the bottom half of the figure signified by the white bars, indicates the negative returns strategies. Taking the average return of all strategies, this research finds that returns from Tweet Terms Trading strategies tested are significantly higher overall than returns from random strategies

($\langle R \rangle_{TTT strategies} = 0.0355, t = 8.705, df = 121, p < 0.001, one sample test$). The t statistic would be calculated as follows:

$$T\text{-statistic} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \quad (5.3)$$

Chapter Five: Text Mining Analysis and Findings

where \bar{x} is the average return of $\langle R \rangle_{TTT \text{ strategies}} = 0.0355$, μ is the mean return of the random strategy $\langle R \rangle_{Random \text{ Strategy}} = 0$, $S=0.0450$ is the standard deviation of the 122 TTT strategies sample and $n= 122$ is the number of the TT trading strategies. Using a one tailed test and 0.001 level of significance and $n-1$ degree of freedom (121 df) the result of t-statistics is $8.705 > 3.1589$ (critical value), which leads to a rejection of the null hypothesis, and concluded that the average returns of TTT strategy is statistically different than the mean return of the uncorrelated random strategy. This result indicates that the TTT strategies proposed here are successful and could produce potential return from implementing them in stock markets. Despite the small average returns of 3.55% of the TTT strategy, these returns exceed the frequently assumed levels of transaction costs for online brokers that range from 0.15%-0.2% where the net profits produced by our TTT strategy are between 3.4%-3.35%. On the other hand, the 'Buy and Hold' strategy resulted in a return of 0.09845 that is a slightly higher return than the overall average of TTT strategies, ($\langle R \rangle_{TTT \text{ strategies}} = 0.0355, t = -15.4791 < 3.1589, p < 0.001, one \text{ sample test}$) which concluded that a 'Buy and Hold' strategy is considered more profitable than the $R_{TTT \text{ strategies}}$. However, considering the performance of the individual term or term combination trading strategy our results show that there are some terms and/or combination of terms trading strategies that outperform the 'buy and hold' strategy. Those strategies are: 'bought', 'yesterday and bought', 'amp and trade', 'appl and trade', 'amp and stock', 'entri', 'chart and price', 'break and support', 'report and intc', 'support', 'report'. In contrast to 'Buy and Hold', the 'Dow Jones' strategy underperformed the average returns of TTT strategies where the 'Dow' strategy resulted in negative returns of -0.0245 compared to 0.0355 of the mean returns of TTT strategies.

Chapter Five: Text Mining Analysis and Findings

Table 5.16: The cumulative average returns and the number of Sell/Buy trades of the Tweet Term Trading TTT Strategies

Term	Cumulative Average Return	Number of Trade		Term	Cumulative Average Return	Number of Trade	
		Buy	Sell			Buy	Sell
report	0.153	136	116	lower+csc	0.061	163	89
support	0.145	135	117	run	0.061	138	114
report+intc	0.129	160	92	post	0.061	136	116
break+support	0.113	142	110	break+look	0.061	143	109
chart+price	0.111	148	104	volume	0.060	131	121
entri	0.110	131	121	move	0.058	124	128
amp+stock	0.109	126	126	sell	0.057	134	118
appl+trade	0.107	131	121	bullish+csc	0.057	171	81
amp+trade	0.107	137	115	bounc	0.056	130	122
yesterday+bought	0.102	175	77	axp+ibm	0.056	149	103
bought	0.101	122	130	head	0.056	136	116
gap	0.098	125	127	cvx+xom	0.055	126	126
day+expect	0.093	153	99	bullish+market	0.054	163	89
cat+call	0.093	135	117	appl+nke	0.054	144	108
bottom	0.089	125	127	qout	0.053	132	120
market	0.088	129	123	current	0.050	136	116
nice	0.084	138	118	axp+look	0.049	189	63
appl+msft	0.081	144	108	csc+amzn+goog	0.049	166	86
watch+follow	0.080	173	79	short	0.047	133	119
cvx+move	0.080	170	82	appl+jpm	0.046	139	113
target	0.076	125	127	spy+msft	0.046	139	113
look+intc	0.074	140	122	buy	0.045	128	124
appl+market	0.073	126	126	csc+break	0.043	166	86
stop	0.071	132	120	amp+daily	0.041	152	100
break+stock	0.071	148	104	goog+sell	0.040	149	103
amp+news	0.067	142	110	look+close	0.039	129	123
bearish	0.067	144	108	amp+head	0.039	150	102
mrk+amp	0.066	143	109	bullish	0.037	135	177
appl+ibm	0.065	138	114	market+time	0.036	143	109
report+low	0.065	151	101	appl+stock	0.036	125	127

Chapter Five: Text Mining Analysis and Findings

Cont'

Term	Cumulative Average Return	Number of Trade		Term	Cumulative Average Return	Number of Trade	
		Buy	Sell			Buy	Sell
look	0.036	130	122	unh+look	0.008	181	71
report+wmt	0.036	171	81	mrk+break	0.006	187	65
cat+break	0.036	145	107	head+move	0.005	178	74
watch+jpm	0.036	148	104	bought+sell	0.005	154	98
stock+current	0.036	137	115	break	0.001	124	128
look+daily	0.035	139	113	appl+wmt	-0.004	131	121
xom+bottom	0.035	199	53	stop+current	-0.006	169	83
day+news	0.033	138	114	cat+run	-0.008	153	99
set	0.033	136	116	sell+pfe	-0.010	184	68
breakout	0.032	136	116	unh+day	-0.013	160	92
cvx+entry	0.031	204	48	stock+amzn+wmt	-0.014	184	68
wmt+Friday	0.031	173	79	flag	-0.017	133	119
stock+amzn	0.028	140	112	chart+post	-0.017	139	113
watch+list	0.027	143	109	strong+support	-0.019	142	110
hold+look	0.027	138	114	hold+gap	-0.019	160	92
hold+play	0.027	138	114	week	-0.020	130	122
lower	0.027	114	138	break+mmm	-0.020	185	67
watch+wmt	0.024	156	96	break+weak	-0.022	136	116
watch+nke	-0.019	152	100	watch+nke	-0.022	148	104
high	0.023	133	119	csc+amzn	-0.022	156	96
chart+flag	0.022	163	89	hold+bounc	-0.023	157	95
unh+gap	0.020	204	48	week+daily	-0.024	134	118
jnj+chart	0.018	161	91	week+time	-0.025	148	104
bull	0.017	129	123	stock+jpm	-0.031	137	115
report+jpm	0.015	152	100	appl+ibm+amzn	-0.033	147	105
unh+hold	0.015	186	66	break+nke	-0.034	163	89
csc+trend	0.013	160	92	cat+mmm	-0.039	149	103
strong	0.013	130	122	amp+dis	-0.042	141	111
jnj+wmt	0.010	141	111	break+bounc	-0.046	163	89
wmt+qout	0.009	152	100	watch+news	-0.051	162	90
				amp+sold	-0.073	144	108
				mcd+trade	-0.079	147	105

Chapter Five: Text Mining Analysis and Findings

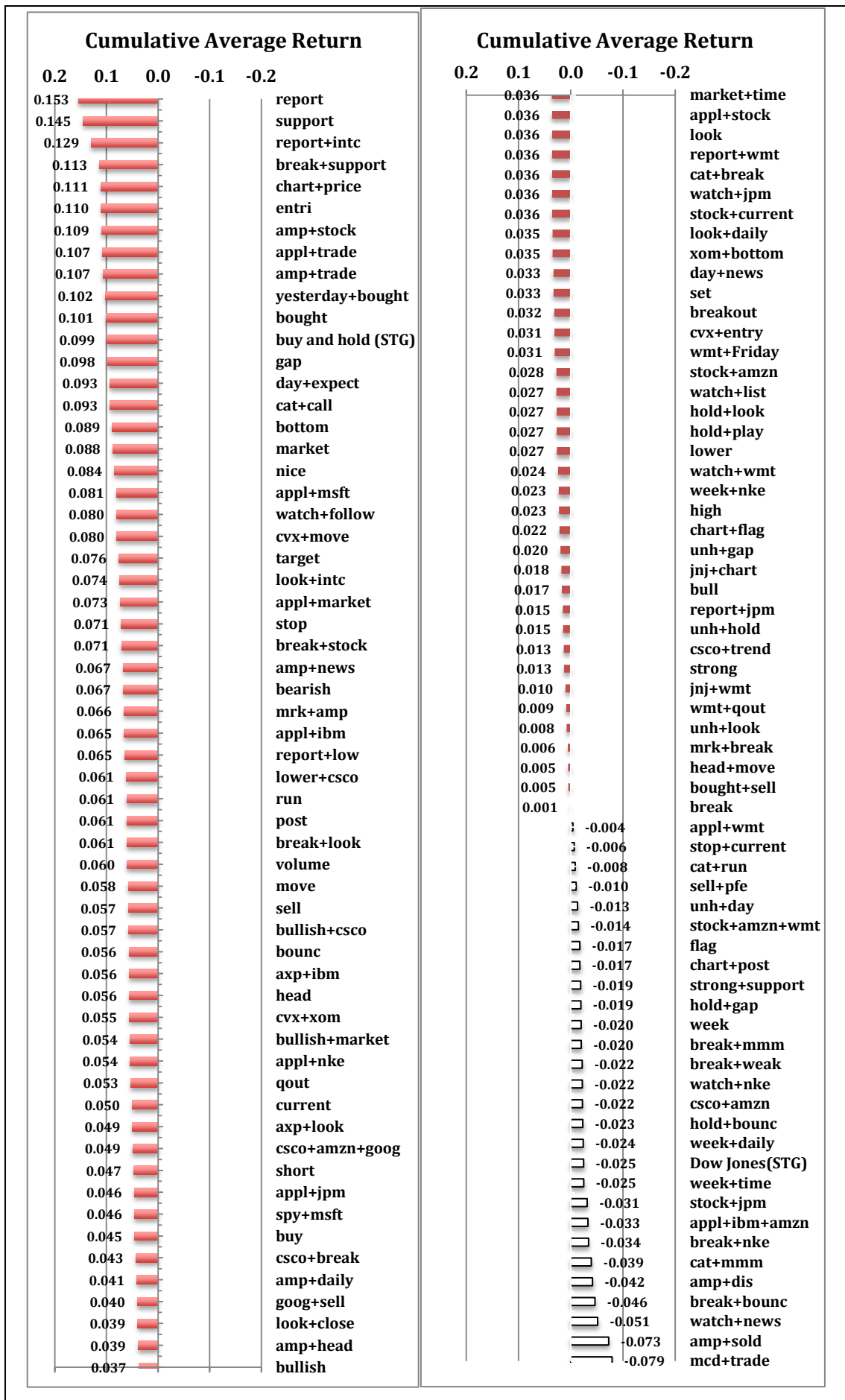


Figure 5.12: Performance of TTT investment strategies based on term related frequency.

Cumulative returns of 122 investment strategies based on tweet term volume frequency are displayed for the entire time period of the study from 3rd of April to 5th of April 2013. Two colors of bars are used to distinguish the positive return strategies from the negative returns. Red bars are used for the positive returns and white bars for the negative returns. The cumulative performance of the “Buy and Hold” strategy and the “Dow Jones” strategy is also shown. Figures depicted next to the bars indicate the returns of a strategy, R , in standard deviation from the mean return of uncorrelated random investment strategy, $\langle R \rangle_{Random Strategy} = 0$. The lines correspond to 0.2, 0.1, 0, -0.1, -0.1 standard deviations of random strategies. All strategies’ returns fall between [0.2, -0.2] standard deviation of RND strategy.

5.9.6 Mean-Variance Analysis

Mean Return should not be the only evaluation factor to consider when evaluating profitability of an investment strategy. A trading strategy is considered superior over another strategy if the risk factor is also involved in the benchmarking process. Mean variance analysis is an element of modern portfolio theory whereby a more efficient investment strategy is made by a rational investor through the process of weighting the variance against expected returns of an asset (Markowitz et al, 2000). Table 5.17 shows the resulting analysis of the mean-variance of each of our studied trading strategies. Note that the Random Strategy is derived from simulations of 1,000 independent realisations of uncorrelated random variables that have a mean of zero and a variance of one whereby at any number of realisations of uncorrelated variables this strategy will always have a mean of zero and a standard deviation of one ($\mu = 0$ and $\sigma = 1$).

Table 5.17 The mean-variance analysis

Mean-Variance Analysis			
Investment Strategy	Mean	Variance	(Mean-variance)
TTT	0.035	0.002	0.033
Random Strategy	0	1	-1
Buy and Hold	0.099	0.558	-0.459
Dow Jones	-0.025	0.559	-0.584

As it can be seen from Table 5.17, our TTT strategies outperform the other benchmark strategies when the risk factor is taken into consideration. All other benchmark strategies (Random, Buy and Hold and Dow Jones Strategy) show a high risk compared to their expected returns indicated by the negative value in the mean-variance column in Table 5.17. While the Buy and Hold strategy showed better performance when the mean returns was the only factor in the evaluation process, it

Chapter Five: Text Mining Analysis and Findings

does not show any good performance when the risks are considered. The TTT strategies are considered the superior strategy among all other benchmark strategies where it exhibits positive returns while maintaining the same level of profitability with a lower level of risk. Although the Buy and Hold strategy is a more profitable investment choice, it however involves much more risk than our TTT strategy.

5.9.7 Portfolio Constructions and Investment Hypothesis

This study aims to investigate the predictability between the TTT decisions obtained from the decision tree algorithm and the market behaviour of stocks of the DJIA index. To start the analysis, three portfolios are constructed namely sell, buy and hold portfolio. Each portfolio consists of all possible terms and/or combination of terms belonging to a particular decision. For example, all sell decision rules extracted from the decision tree corresponding to the sell class will be listed under sell portfolio. The same with the buy and hold portfolios, where all the decision rules belonging to the buy or hold class will constitute the buy and hold portfolios respectively. Table 5.18 shows the list of terms constituting the sell, buy and hold portfolios.

Table 5.18: The term trading strategies in the sell, hold and buy portfolios

Portfolio	Term Trading Strategies
Sell Portfolio	“bearish”, “bottom”, “bounce”, “flag”, “gap”, “lower”, “sell”, “short”, “stop”, “support”, “volume”, “unh+gap”, “mrk+amp”, “mrk+break”, “report+intc”, “report+wmt”, “break+mmm”, “break+nke”, “xom+bottom”, “amp+head”, “look+intc”, “stock+jpm”, “week+nke”, “wmt+qout”, “break+bounc”, “break+stock”, “break+support”, “break+weak”, “bullish+market”, “chart+flag”, “chart+price”, “day+expect”, “day+news”, “head+move”, “hold+gap”, “hold+look”, “look+close”, “look+daily”, “stock+current”, “strong+support”, “support+break”, “week+daily”, “week+time”, “yesterday+bought”
Hold Portfolio	“current”, “entri”, “market”, “post”, “qout”, “report”, “set”, “week”, “appl+ibm+amzn”, “csc+amzn+goog”, “appl+jpm”, “appl+stock”, “appl+wmt”, “appl+market”, “appl+trade”, “appl+msft”, “goog+sell”, “sell+pfe”, “watch+jpm”, “watch+nke”, “watch+wmt”, “mcd+trade”, “stock+amzn”, “jnj+wmt”, “wmt+friday”, “chart+post”, “market+time”, “watch+follow”, “watch+list”, “watch+news”,
Buy Portfolio	“bought”, “break”, “breakout”, “bull”, “bullish”, “buy”, “head”, “high”, “look”, “move”, “nice”, “run”, “strong”, “target”, “lower+csc”, “bullish+csc”, “csc+amzn”, “csc+trend”, “csc+break”, “cvx+move”, “cvx+entry”, “cvx+xom”, “cat+mmm”, “cat+break”, “cat+run”, “cat+call”, “unh+hold”, “unh+day”, “unh+look”, “appl+nke”, “report+jpm”, “appl+ibm”, “spy+msft”, “axp+ibm”, “axp+look”, “amp+news”, “amp+stock”, “amp+daily”, “amp+sold”, “amp+trade”, “amp+dis”, “stock+amzn+wmt”, “jnj+chart”, “bought+sell”, “break+look”, “hold+bounc”, “hold+ply”, “report+low”, “stop+current”

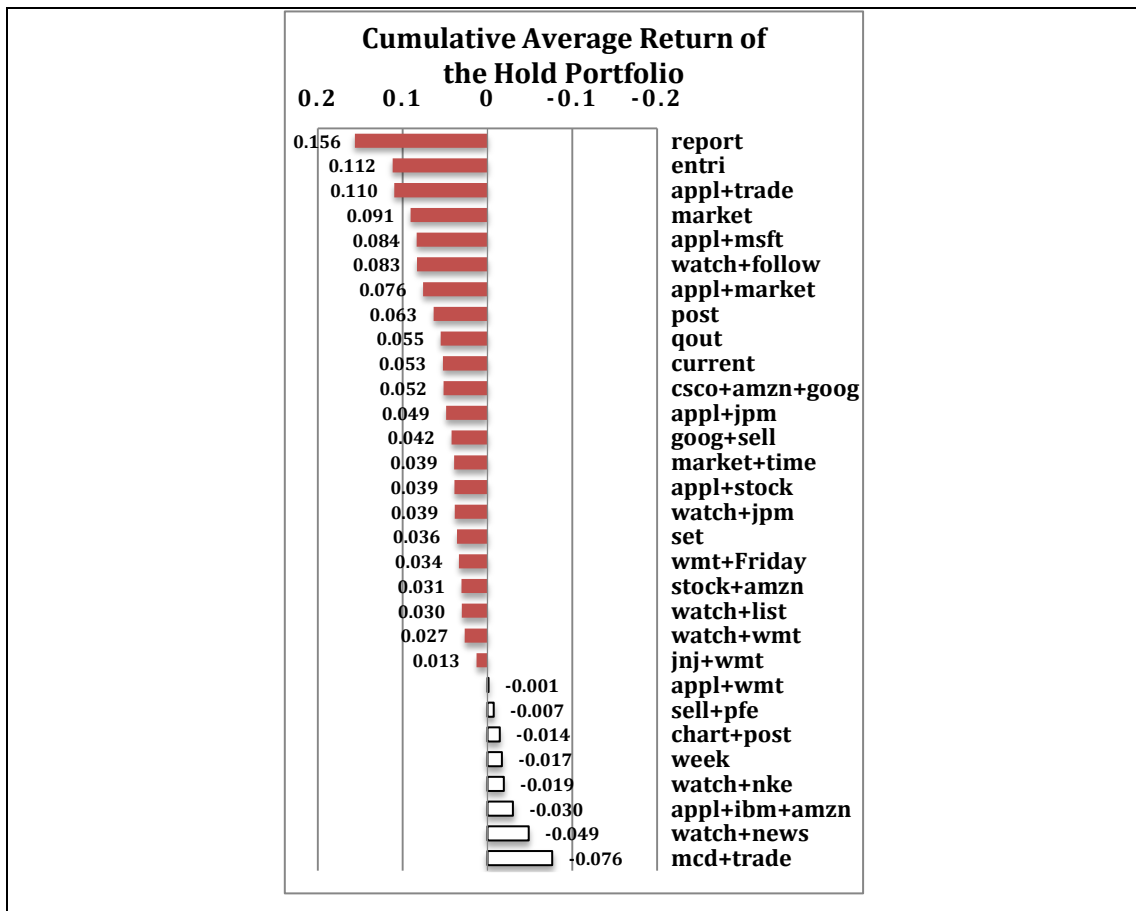
As it can be seen from Table 5.18, a total of 122 trading decisions were

Chapter Five: Text Mining Analysis and Findings

returned from the decision tree algorithm C4.5. The sell portfolio consists of 49 terms, while 44 terms indicated buying decisions and 30 terms represented holding decisions.

5.9.8 Cumulative Performance of the Sell, Buy and Hold Portfolios

This section documents the strategies' returns of the portfolio constructed in the previous section. The returns of all terms constituting each portfolio are calculated based on the trading strategy described earlier. Figure 5.13 shows the average returns of the 122 different terms distributed based on their trading decisions in the sell, buy and hold portfolio. The most successful strategies are those terms composing the sell portfolios that yielded higher average returns of 0.0408 compared to 0.0369 and 0.0366 for the sell, buy and hold portfolio respectively. All portfolios returns are statistically significant and higher overall than returns of the random investment strategy. The individual t-statistics of each portfolio are sufficiently large to be significant to reject the null hypothesis that the mean portfolio returns are equal to the mean return of the random strategy.



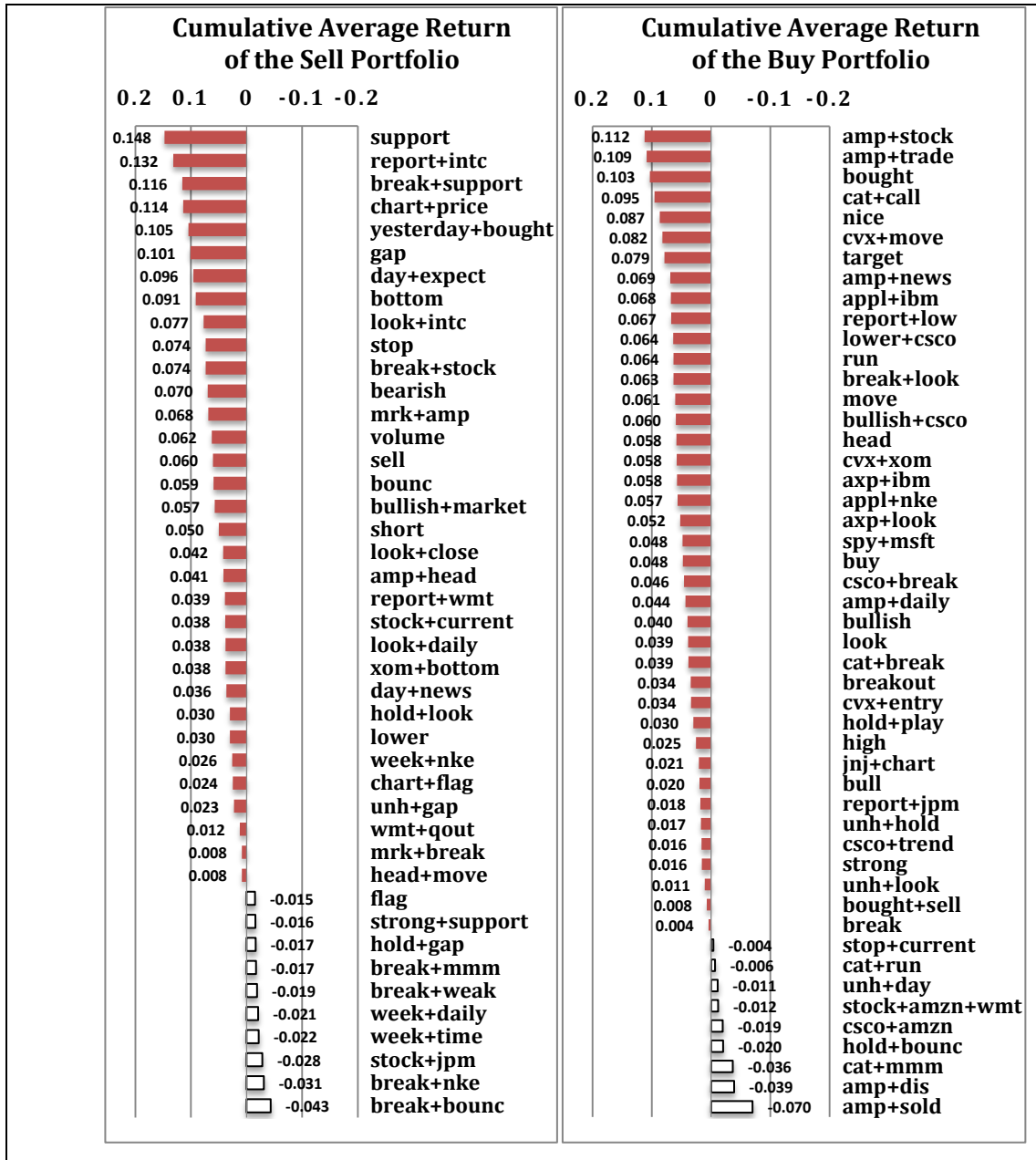


Figure 5.13: Performance of sell, buy and hold portfolios strategies.

Cumulative returns of 122 investment strategies distributed based on their trading decision into the sell (43 terms), buy (49 terms) and hold (30 terms) portfolios. Two colors of bars are used to distinguish the positive return strategies from the negative returns. Red bars are used for the positive returns and white bars for the negative returns. The cumulative performance of the “Buy and Hold” strategy and the “Dow Jones” strategy is also shown. Figures depicted next to the bars indicate the returns of a strategy, R , in standard deviation from the mean return of uncorrelated random investment strategy, $\langle R \rangle_{Random Strategy} = 0$. The lines correspond to 0.2, 0.1, 0, -0.1, -0.1 standard deviations of random strategies. All strategies’ returns fall between [0.2, -0.2] standard deviation of RND strategy. The average returns of all of our portfolios (sell, buy and hold) are positive. The t-statistics of the portfolios’ returns using one tailed test are ($\langle R \rangle_{sell portfolio} = 0.0408, t = 5.600 > 3.2959 df = 42, p < 0.001$); ($\langle R \rangle_{buy portfolio} = 0.0369, t = 6.506 > 3.2689 df = 48 p < 0.001$); ($\langle R \rangle_{hold portfolio} = 0.0366, t = 3.997 > 3.3969 df = 29, p < 0.001$) for the sell, buy and hold portfolio respectively.

5.9.9 Investment Hypothesis Evaluation

This section evaluates the effectiveness of the trading strategies in anticipating subsequent moves in financial markets. The results show that performance of the Tweet Term Trading TTT strategies varies with cross terms or (combination of terms) that appeared in tweet postings. This study additionally found that the different buy, sell and hold portfolios produce different average cumulative returns suggesting that each of these portfolios would have different roles in affecting our strategy returns. The empirical result of this research is implemented based on a two-part investment hypothesis. The two parts of this hypothesis are:

- Increases in the prices of the DJIA were preceded by a decrease in the volume frequency of related term, which prompts one to sell or take a short position.
- Decreases in the prices of the DJIA were preceded by an increase in the volume frequency of related terms, which prompts one to buy or take a long position.

It is therefore important to test and verify these two strategy components. To validate the significant role each part of this hypothesis plays, these two strategy components are implemented by examining the asymmetric effects of the increase and decrease of the mean relative change in the tweet term frequency. At each day t the mean relative change in the term frequency for the sell, buy and hold portfolios over the previous five days average is calculated, as follows:

$$x_{it} = \Delta n(t, \Delta t) / N(t-1, \Delta t) \quad (5.4)$$

where x_{it} is the mean relative change of term i in a portfolio at a time t , $n(t)$ is the volume frequency of a term appeared in a given day and $N(t-1, \Delta t) = (n(t-1) + n(t-2) + \dots + n(t-\Delta t)) / \Delta t$ is the average number of term frequency of the previous 5 days.

In order to test each part of our hypothesis, it would be expected that the sell portfolio terms would confirm the first part, in which the appearance of such terms signify a sell signal in the stock market (short position), while the buy portfolio terms would be used to explain and verify the second part of the investment hypothesis, fuelling the fact that the appearance of those terms in tweet messages indicates a buy signal to other market participants (long position). Whereas, the holding decision

Chapter Five: Text Mining Analysis and Findings

would have a limited effect on the profitability position of an investor in a capital market and one would expect that the returns of the hold portfolio may have equal feedback to the effect of the increases and decreases of the mean relative frequencies of the tweet terms. The study formally investigates whether the language of StockTwits provides new information about investment decisions in stock markets and whether stock market prices efficiently incorporate this information. This approach also allows exploring relationships between the magnitudes of the increases and decreases in volume frequency of the related term and the magnitude of the subsequent returns of our trading strategies.

To isolate the effects of an increase or decrease in the mean relative change of a term, the following indicator variables are computed,

$$I^+ = \begin{cases} 1 & \text{if } x_{it} > 0 \\ 0 & \text{otherwise} \end{cases} \quad (5.5a)$$

$$I^- = \begin{cases} 1 & \text{if } x_{it} < 0 \\ 0 & \text{otherwise} \end{cases} \quad (5.5b)$$

For the increase in the mean relative frequency the indicator variable I^+ takes the value of one when x_{it} is positive, and the value of zero otherwise. Likewise, for the decrease in the mean relative frequency the indicator variable takes the value of one when x_{it} is negative, and the value of zero otherwise. Accordingly, those two variables for each term undertaken in this analysis are created.

The focus is on Ordinary Least Square (OLS) regression estimates of the effect of increases and decreases of the mean relative frequency of term of different portfolios on the subsequent returns of our investment strategy relative to the occurrence of the terms in StockTwits postings. Therefore in this section, panel regression with cross section fixed effect for each term i is employed to estimate the contemporaneous regressions for each portfolio j (sell, buy and hold) separately. Regressions will be estimated using standard ordinary least squares (OLS) techniques, where the return from the proposed trading strategy is treated as a dependent variable and regressed on two independent variables; the increase in the mean relative change of the term frequency indicated by x_{it}^+ , and x_{it}^- indicates the decreases in the mean relative change of the term frequency. The market return of DJIA index return (MKT_t) is added in the regressions as a control variable to control for overall market wide effects. The OLS subsequent return regression equations for each of the three

Chapter Five: Text Mining Analysis and Findings

portfolios are shown in Table 5.19 and can be expressed as:

$$R_{it, Sell Portfolio} = \alpha_1 + \beta_1^+ x_{it}^+ + \beta_2^- x_{it}^- + \beta_3 MKT_t + \beta_4 Dummy + \varepsilon_{it}, \quad (5.6)$$

$$R_{it, Buy Portfolio} = \alpha_2 + \beta_5^+ x_{it}^+ + \beta_6^- x_{it}^- + \beta_7 MKT_t + \beta_8 Dummy + \varepsilon_{it}, \quad (5.7)$$

$$R_{it, Hold Portfolio} = \alpha_3 + \beta_9^+ x_{it}^+ + \beta_{10}^- x_{it}^- + \beta_{11} MKT_t + \beta_{12} Dummy + \varepsilon_{it}, \quad (5.8)$$

The OLS estimates of the coefficients β_s in Eqs. (5.6), (5.7) and (5.8) are the primary focus of these regression equations. These coefficients describe the dependence of the positive (increase) and negative (decrease) variation in mean relative change of volume of a term that appeared in a tweet message on the subsequent change of returns (returns of our investment strategy calculated in an earlier section). Table 5.19 summarises the estimates of β_s .

Table 5.19: Predicting Portfolio's Trading Strategy Returns Based on the Asymmetric Effects of the Increase and Decrease in the Mean Relative Changes of the Term Related Frequencies.

On data measured on daily frequency, panel regressions with term fixed effects are estimated separately for each portfolio j = (Sell, Buy and Hold) where trading strategy returns are used as a dependent variable. The independent variables were obtained from the mean relative change in volume of a term appeared in StockTwits postings in a particular day (t): The positive (increase) x_t^+ and negative (decrease) x_t^- variation in mean relative change of volume of a term (i) in portfolio (j). This table shows the predictive power of the positive and negative variation in tweet term volume in explaining the subsequent change of trading strategy returns of different portfolios. In all regressions, Market return is added as a control variable. Market return denotes the log difference of DJIA price. To control for Monday return anomaly, dummy variable for first day of the week is added in all portfolio returns regressions.

Subsequent Return $R_{(t+1)}$	Increase in Mean RCHG x_{it}^+	Decrease in Mean RCHG x_{it}^-	Market	Dummy	R^2	Durbin Watson
Sell Portfolio	0.0060 (1.5421)	0.0324** (2.1687)	-0.0185* (-1.9185)	0.0026 (0.1514)	0.527	2.021
Buy Portfolio	0.0052** (1.9915)	0.00415 (0.3064)	-0.0211** (-2.3420)	-0.0108 (-0.6719)	0.353	2.017
Hold Portfolio	0.0087** (1.9919)	0.0371* (1.9073)	-0.0215* (-1.8563)	-0.0171 (0.3927)	0.623	2.016

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$, t-statistics in parenthesis below the coefficients.

Chapter Five: Text Mining Analysis and Findings

The regression results of the return equations of the sell, buy and hold portfolios as shown in Table 5.19, are largely as one would expect. That is, that the term trading strategies constituting each portfolio were generating positive returns indicated by the positive β_s coefficients (regardless of not being statistically significant) of the asymmetric effects of both the increase and decrease in the mean relative change of the terms frequency in all three portfolios regressions. However, to test the two parts of the investment hypothesis this research needs to, therefore, investigate in depth the effect of the increase and decrease in the mean relative changes of the terms frequency in each portfolio separately. The sell portfolio regression found a statistical significant coefficient of the decrease in the mean relative change in the term frequency ($\beta_2 = +0.0324$, p value < 0.05) while an increase on the other hand, exerted no statistical significance in forecasting the portfolio trading strategy returns. This suggests that the decrease in the mean relative change of the sell terms that appeared in tweet postings have a proportionally larger impact on the subsequent returns of the TTT strategies of DJIA index rather than the effect of an increase.

Since the appearance of the terms in the sell portfolio signifies a sell signal to market participants, a decrease in the appearance of such terms conveys a good signal before market rises. One possible explanation of these results could be interpreted from a psychological viewpoint. The most common words that are more likely to appear in sell messages in StockTwits are negative words like “break” and “lower”, “bottom” as well as words like; “sell”, “bearish” and “short” which give a clear sign indicating that investors expected the discussed stocks to fall. Therefore, a decreased appearance of such negative words/terms is an indication of a decrease in an investor’s bearishness, which implies good signals to their relative peers in the market that prices will start to recover and move upwards. These findings strengthen the first part of the trading hypothesis that an increase in DJIA prices were preceded by a decrease in the volume frequencies of the sell terms, which prompts one to sell or take short position.

The buy portfolio regression however shows inverse results to what was found in the sale portfolio regression. There is no statistical significant effect in the relation between the decrease in the mean relative change of the related term frequency and the buy portfolio returns. But it has been found that the increase in the mean relative

Chapter Five: Text Mining Analysis and Findings

change of the term frequency exerts a statistically significant influence in forecasting the buy portfolio trading strategies returns of DJIA index. Despite the statistical significant effect of the increase in the mean relative change, the estimated effect of +0.0052 is very small in magnitude. However, even here such tiny price effects would be difficult to take advantage of because this potential gain would likely to be offset even by transaction costs resulting in relatively trivial gain if not negative. Hence, an increase in the mean relative change of the buy terms is more likely to be followed by a decrease in DJIA prices where people see a buying opportunity and tend to take a long position in the market. Since, the buy terms that appeared in StockTwits messages indicates an investor's optimism and provides a "buy" signal to the market participants, an increase in such terms will increase bullishness of investors where they are more likely to see a buying opportunity of stocks expecting prices to fall. Our evidence supports the "bargain shopper" hypothesis: the market speculators, who see stocks becoming a bargain, see a buying opportunity and become bullish (Brown and Cliff, 2004). These results however, confirm the second part of our trading hypothesis that decreases in the prices of the DJIA were preceded by an increase in the volume frequency of related term, which prompts one to buy or take the long position in capital markets.

Looking at the hold portfolio regression in Table 5.19, it can be noticed that both the increase and decrease in the mean relative change of term frequency exerts a statistically significant positive effect in explaining our strategy returns indicated by the significant coefficients of $\beta_9^+ = + 0.0087$, p-value < 0.1 and $\beta_{10}^- = + 0.0371$, p-value < 0.05 for the increase and decrease in the mean relative change respectively. The estimated coefficients of both effects are economically small, which is in the context of our investment hypothesis; H_a : an increase in DJIA prices preceded by a decrease in the term volume frequency (which recommends investors to sell and take short position). This will be offset by the inverse effect of the second part of the hypothesis; H_b : a decrease in DJIA prices preceded by an increase in the term related frequency (which prompts investors to buy and take the long position). This result is not surprising, however, where in real life economics a holding decision has taken place where an investor is not optimistic enough to buy a stock, but not pessimistic enough to sell a stock. This is also true if one gets closer to investigate the nature of the words/terms comprising the hold portfolio, where an equal balance of positive and

Chapter Five: Text Mining Analysis and Findings

negative terms/combinations of terms are more likely to be found. It also contains neutral words like “report”, “qout”, “entri” as well as the name of the companies like; “cat”, “jpm” and “wmt”. The appearance of these kinds of terms in tweet messages would cause an investor to hold a neutral opinion about particular traded stocks where they most probably take holding decisions rather than buy or sell. The coefficients β_3 , β_7 and β_{11} of the market return (DJIA) index were statistically and negatively significant in all portfolios’ regressions whereas, the dummy variable of the first day of the week effect indicated by the coefficients β_4 , β_8 and β_{12} , reported insignificant in all regression equations in Table 5.19.

5.9.10 Discussion

This study proposes a novel approach by combining text mining, feature selection and a decision tree model to quantify and predict investor sentiment from a stock micro-blogging forum (StockTwits) of DJIA companies. The experiments reported in this chapter provide quantifications of the StockTwits semantic terms trading decisions extracted from the decision tree algorithm, while providing a linkage between changes in the volume of semantic terms and subsequent stock market moves. The findings of this research proved the success of the investment-trading hypothesis implemented for the different semantic terms trading strategies of StockTwits. This research suggests two subsequent stages in the decision making process of investors using both StockTwits semantic terms and stock market data. Trends to sell short a stock at higher prices resulted from a decrease in the volume appearance of negative words (terms constituting the sell portfolio) in the tweet postings, while the trends to buy or take long positions resulted from an increase in the volume appearance of positive words (terms constituting the buy portfolio) in tweet postings.

Overall, our results indicate the existence of the asymmetric effect of StockTwits sentiments indicated by the (sell, buy hold) portfolios on the subsequent moves in the stock market. This study confirms that StockTwits postings contain valuable information and precede trading activities in capital markets. Changes in the average occurrences of different semantic terms in StockTwits postings informed decisions on whether to buy or sell the DJIA stocks. These findings may yield

Chapter Five: Text Mining Analysis and Findings

promising insights into the potential provision of an investment support mechanism for analysts, investors and their peers. Practically, this could be used to determine the precise time when stocks are to be held, added (buy) or removed (sell) from a portfolio, thus yielding the maximum return on the investment for the investor. This could save time and effort and will lead to making a better-informed investment decision in the capital market.

5.10 Chapter Summary

This chapter presented the analysis and findings of the text-mining procedure for extracting and predicting sentiments from stock-related micro-blogging data. A comprehensive textual analysis of each of the three machine-learning algorithms selected in this empirical fieldwork was evaluated and discussed for the purpose of selecting the most accurate text-mining techniques for predicting sentiment analysis on StockTwits. An essential data analysis tool of a text-mining technique called feature selection method was performed to select a set of relevant features from datasets based on some predetermined criteria. Wrapper and filter approaches are the two approaches of feature selection methods that are used to extract the most relevant features as both provide greater insights into the relevancy of the data being used.

This chapter also presented two exciting applications of both approaches of feature selection methods (filter and wrapper), as these applications offer practical implications for market participants in the capital market. The aim of these applications is to provide real-time investment ideas that may provide investors and their peers with an investment decision support mechanism. Based on the automated classifications of the selected classifier, the overall categorisation of StockTwits data is provided to determine the proportion of the tweet messages that should be classified into buy, hold or sell classes. These findings are then aggregated to calculate the StockTwits variables for further analysis in the forthcoming chapters, which will then be correlated with other financial market indicators. The forthcoming chapters will provide detailed investigations of the relationship between the StockTwits features and financial market to determine whether or not the sentiments extracted from stock micro-blogging messages might explain the stock market dynamics and behaviour in the capital market.

CHAPTER SIX: EMPIRICAL FINANCE ANALYSIS AND DISCUSSION

6.1 Introduction

This research aims to gain a deeper understanding and undertake a more holistic analysis of the predictive power of stock micro-blogging sentiments in predicting stock market behaviour in the financial market.

After presenting the findings from the first stage of the analysis and providing a detailed discussion of the sentiment analysis in the previous chapter, this chapter presents the findings of the empirical finance analysis in order to test the hypotheses and to develop a solid and more comprehensive understanding of the impact of stock micro-blogging sentiment on forecasting stock market movements. The empirical finance analysis confronts economic theories (e.g. efficient market hypothesis, random walk and behavioural finance) about stock prices and other financial market indicators (e.g. returns) with real-world finance data. These types of analyses deal with how prices and returns should behave in an efficient market through the applications of different econometric modelling and various statistical techniques to assess whether the data under study support such behaviour. Theoretical discussions are provided under each part of the findings whereby the key results from the findings are linked to existing theoretical perspectives to confirm or reject the formulated hypotheses of this research study.

This Chapter consists of seven sections including this introduction. Section 6.2 provides a brief summary of descriptive statistics of the data used in this research. Section 6.3 shows an initial analysis of the distributions of the tweet posting volumes by company, time and day of the week. Section 6.4 describes the contemporaneous correlations (the pairwise correlation and contemporaneous regressions) between StockTwits features and financial market indicators. The lead-lag relationships between stock micro-blogging features and financial market indicators employing Vector Auto Regressive (VAR) models are investigated in Section 6.5. The VAR models highlight how the three features of StockTwits (i.e. bullishness, message volume and agreement) affect financial market variables (i.e. returns, volatility and

Chapter Six: Empirical Finance Analysis and Discussion

trading volumes), which may cause movement in the financial market and vice versa, using various statistical methods and econometric techniques. Section 6.6 presents the impulse response function of the three VAR models (VAR: returns, volatility and trading volume). Section 6.7 summarises this chapter.

6.2 Descriptive Statistics and Preliminary Analysis

This Section presents the summary statistics for DJIA companies during the period between 3rd of April 2012 and 5th of April 2013. The sample includes 7,560 observations for the 30 companies making up the DJIA over a period of 252 days. Returns were calculated as the log difference in prices²², daily trading volumes were calculated as the natural logarithm of the shares outstanding, and following Garman and Klass (1980) in estimating the daily volatility based upon the historical opening, closing, high, and low prices.²³ Bullishness is the proxy for investor sentiment of a particular message, message volume represents the number of daily messages per company, and level of agreement measures the concurrence of messages of a particular company with respect to their sentiment (i.e. buy vs. sell messages). The formulas used to compute all these variables were reported and discussed in more detail in the methodology chapter of this thesis. For simplicity and greater readability, returns are scaled by 100 while volatility is scaled by 10,000. Table 6.1 reports the descriptive statistics of the above-mentioned variables.

Table 6.1: Summary Statistics by Variable

Variables	Mean	Std. Division	Minimum	Maximum
<u>Market Features</u>				
Return	0.0426	1.1920	-10.9630	10.4931
Volatility	0.0004	1.15686	-17.6466	15.3558
Trading Volume	14.5977	0.9174	10.8438	17.9142
<u>Tweet Features</u>				
Bullishness	0.3084	0.1616	0.0000	1.0000
Message Volume	2.5132	1.4388	0.0000	8.0465
Agreement	0.1165	0.2284	0.0000	1.0000

The average return at which our sample of stocks traded during the period under study was \$4.26. The highest return observed in the data was \$10.49 and the

²² Winsorisation of return has been performed to reduce the effect of outlier values in the data. However, the reported results do not change the significance of the relationship of returns with other studied variables.

²³As noted previously in the methodology chapter, because of the volatility persistence, the changes in volatility are used throughout the analysis of this thesis instead of volatility levels.

Chapter Six: Empirical Finance Analysis and Discussion

lowest return observed was \$-10.96. This range is unusually large for firms that are included in the DJIA index. The trading volumes of DJIA have a high standard deviation of approximately 91.7%, suggesting that the levels of traded shares are highly volatile during the sample period. Figure 6.1 depicts the monthly average movement of each of the six variables employed in this study.



Chapter Six: Empirical Finance Analysis and Discussion



Figure 6.1: The monthly average movement of StockTwits variables (bullishness, message volume and agreement) and financial market variables (trading volume, returns and volatility).

*Note that the change in volatility is also depicted in addition to volatility level since the former will be used in all the analysis of this thesis.

Chapter Six: Empirical Finance Analysis and Discussion

Descriptive statistics for all companies in our sample are also reported in Table 6.2. As it can be seen from the table, there are more messages posted for all firms in general and the messages are often more bullish. The number of messages ranges from 1,312 messages on United Technologies Corporation (\$UTX) to 35,336 messages on JP Morgan (\$JPM).

Table 6.2: Summary Statistics by Company

Ticker	Company Name	Bullishness	Total Message Volume	Agreement	Return ^a	Volatility ^b	Trading Volume
Axp	American Express Company	0.369	3,165	0.141	0.058	0.96	1,271,870
BA	The Boeing Company	0.185	4,867	0.208	0.063	0.99	995,711
CAT	Caterpillar Inc	0.498	18,020	0.079	-0.143	1.68	1,371,349
CSCO	Cisco Systems, Inc	0.090	11,512	0.120	-0.002	1.38	9,551,762
CVX	Chevron Corporation	0.449	4,564	0.111	0.033	0.75	1,706,137
DD	E. I. du Pont de Nemours and Company	0.329	1,564	0.148	-0.038	0.94	1,188,046
DIS	The Walt Disney Company	0.418	5,948	0.141	0.073	0.86	2,055,812
GE	General Electric	0.338	6,728	0.095	0.037	0.92	8,266,652
GS	The Goldman Sachs Group, Inc	0.328	28,972	0.058	0.104	1.65	935,033
HD	The Home Depot	0.214	6,924	0.134	0.017	0.93	1,763,522
IBM	International Business Machines Corporation (IBM)	0.493	17,372	0.080	-0.070	0.62	936,758
INTC	Intel Corporation	0.270	16,608	0.091	0.102	1.32	9,712,067
JNJ	Johnson & Johnson	0.336	4,988	0.134	0.025	0.35	2,616,404
JPM	JPMorgan Chase & Co.	0.246	35,336	0.064	0.001	1.77	5,163,987
KO	Coca-Cola	0.213	5,716	0.115	-0.026	0.56	4,010,363
MCD	McDonald's	0.373	7,952	0.108	0.077	0.51	1,455,624
MM	Minnesota Mining and Manufacturing Company (3M)	0.166	1,736	0.181	0.035	0.59	800,930
MRK	Merck & Co., Inc.	0.319	3,160	0.143	-0.070	0.69	2,761,368
MSFT	Microsoft Corporation	0.238	33,700	0.072	0.039	1.05	11,664,016
NKE	Nike, Inc.	0.103	10,620	0.104	0.134	1.26	1,244,767
PFE	Pfizer, Inc	0.386	4,460	0.139	0.115	0.73	6,213,640
PG	Procter & Gamble Co.	0.458	3,228	0.158	0.089	0.45	2,288,327
T	American Telephone & Telegraph (AT&T)	0.096	9,832	0.132	0.076	0.73	4,962,166
TRV	The Travelers Companies	0.467	1,344	0.122	0.032	0.68	617,959
UNH	UnitedHealth Group Inc.	0.180	2,636	0.109	0.023	1.51	1,369,829
UTX	United Technologies Corporation (UTC)	0.258	1,312	0.118	0.090	0.97	886,375
V	Visa Inc.	0.295	11,224	0.095	0.085	1.12	693,900
VZ	Verizon Communications	0.463	6,892	0.107	0.108	0.76	2,672,378
WMT	Wal-Mart Stores, Inc.	0.206	10,740	0.096	0.047	0.62	1,923,928
XOM	Exxon Mobil Corp	0.468	7,904	0.094	0.058	0.63	3,791,582

Note: a Return is the average returns of the all trading days over the sample period of study. b average of the daily volatility measures as explained in chapter 4 are scaled by 10,000 for easier readability.

Chapter Six: Empirical Finance Analysis and Discussion

6.3 Distribution of StockTwits Postings

6.3.1 Distribution of StockTwits by DJIA Tickers

As noted in the previous chapter, StockTwits data are collected from the StockTwits website Application Programming Interface (API) and are used as the main data source to conduct this research study. StockTwits postings were pre-processed and those posts without any ticker, with more than one ticker and those not in the DJIA index were removed, leaving 289,024 valid postings consisting of 30 stock tickers; 27 in NYSE (contributing about 227,194 of the total postings) and three in NASDAQ (contributing about 61,830 of the total postings), containing the dollar-tagged ticker symbol of the 30 stock tickers of DJIA (Dow 30). Table 6.3 presents the list of the companies constituting the DJIA index along with their ticker symbols, the related volume of postings for each company and the traded stock exchange in which each of these companies is listed.

Table 6.3: The list of the DJIA Index stock tickers

Stock Ticker	Name of the Company	Number of Posts	Stock Market
\$AXP	American Express Company	3165	NYSE
\$BA	The Boeing Company	4867	NYSE
\$CAT	Caterpillar Inc	18020	NYSE
\$CSCO	Cisco Systems, Inc	11512	NASDAQ
\$CVX	Chevron Corporation	4564	NYSE
\$DD	E. I. du Pont de Nemours and Company	1564	NYSE
\$DIS	The Walt Disney Company	5948	NYSE
\$GE	General Electric	6728	NYSE
\$GS	The Goldman Sachs Group, Inc	28972	NYSE
\$HD	The Home Depot	6924	NYSE
\$IBM	International Business Machines Corporation (IBM)	17372	NYSE
\$INTC	Intel Corporation	16608	NASDAQ
\$JNJ	Johnson & Johnson	4988	NYSE
\$JPM	JPMorgan Chase & Co.	35336	NYSE
\$KO	Coca-Cola	5716	NYSE
\$MCD	McDonald's	7952	NYSE
\$MMM	Minnesota Mining and Manufacturing Company (3M)	1736	NYSE
\$MRK	Merck & Co., Inc.	3160	NYSE
\$MSFT	Microsoft Corporation	33700	NASDAQ
\$NKE	Nike, Inc.	10620	NYSE
\$PFE	Pfizer, Inc	4460	NYSE
\$PG	Procter & Gamble Co.	3228	NYSE
\$T	American Telephone & Telegraph (AT&T)	9832	NYSE
\$TRV	The Travelers Companies	1344	NYSE
\$UNH	UnitedHealth Group Inc.	2636	NYSE
\$UTX	United Technologies Corporation (UTC)	1312	NYSE
\$V	Visa Inc.	11224	NYSE
\$VZ	Verizon Communications	6892	NYSE
\$WMT	Wal-Mart Stores, Inc.	10740	NYSE
\$XOM	Exxon Mobil Corp	7904	NYSE
Total		289,024	

Chapter Six: Empirical Finance Analysis and Discussion

A list of the top ten stock tickers with a corresponding number of postings is shown in Table 6.4. Excitingly, the top ten stock tickers of the DJIA account for approximately 67% of all postings. These stocks are the most popular stocks on the Dow Index and the companies are heavily and favourably discussed in StockTwits by investors, analysts and other market professionals. This finding concurs with that of Huberman (2001), who provided evidence that people tend to invest in familiar stocks - for instance, their own company's stocks, stocks of the firms they know and that are visible in the investors' lives, or the stocks that are discussed more intensively in the media - while often ignoring the fundamental principles of diversification and portfolio theory.

Table 6.4: Distribution of postings by top ten tickers

Rank	Stock Ticker	Stock Market	Total
1	\$JPM	NYSE	35,336
2	\$MSFT	NASDAQ	33,700
3	\$GS	NYSE	28,972
4	\$CAT	NYSE	18,020
5	\$IBM	NYSE	17,372
6	\$INTC	NASDAQ	16,608
7	\$CSCO	NASDAQ	11,512
8	\$V	NYSE	11,224
9	\$WMT	NYSE	10,740
10	\$NKE	NYSE	10,620
Total			194,104

6.3.2 Distributions of StockTwits by time of day

Since this research study centred on the U.S. market, as the DJIA index was chosen as the context of this study, it is very important to align StockTwits messages with U.S. market timing hours. As the DJIA index is traded in NYSE and NASDAQ, these markets are open from 9:30am to 4:00pm Eastern time. It is believed that the market and its participants behave somewhat differently at the times when the market opens and closes from the way they behave most of the day. As noted by Bacidore and Lipson (2001), the trading procedures at the market opening and closing times on the NYSE are different from trading during the rest of the day.

In line with Antweiler and Frank (2004b), messages are aligned with US

Chapter Six: Empirical Finance Analysis and Discussion

market hours, where the messages posted after 4:00 pm (the time the market closes) are combined with pre-market messages at 9:30 am (the time the market opens) on the next trading day. There are a number of reasons for this time alignment. The primary reason is that the effect of these messages on the market indicators can only appear on the next trading day. Second, the market behaves differently over different time spans during the day. For example, the trading activity tends to be different at the open and at the close of trading from trading during the rest of the day. Third, the availability of various traders in the market along with their self-motivation to trade over different time periods will have different effects on the market over that time. Small traders are most likely to trade more during the evening as they tend to be busy at work in the morning. Sometimes they may contact their brokers early in the morning, which results in more trading activity just before the market opens. On the other hand, many institutional investors will want to avoid overnight risk by closing out their portfolio positions at the end of the day. Figure 6.2 shows the distribution of the tweet messages during the day. Consistent with previous studies, it is observed that a high volume of tweets are posted during working hours, typically between 10:00am and 5:00pm while the market is open. This is analogous with Antweiler and Frank's (2004b) finding that the market auction takes approximately half an hour after opening; hence, the effect of tweets appears at 10:00 am instead of 9:30am.

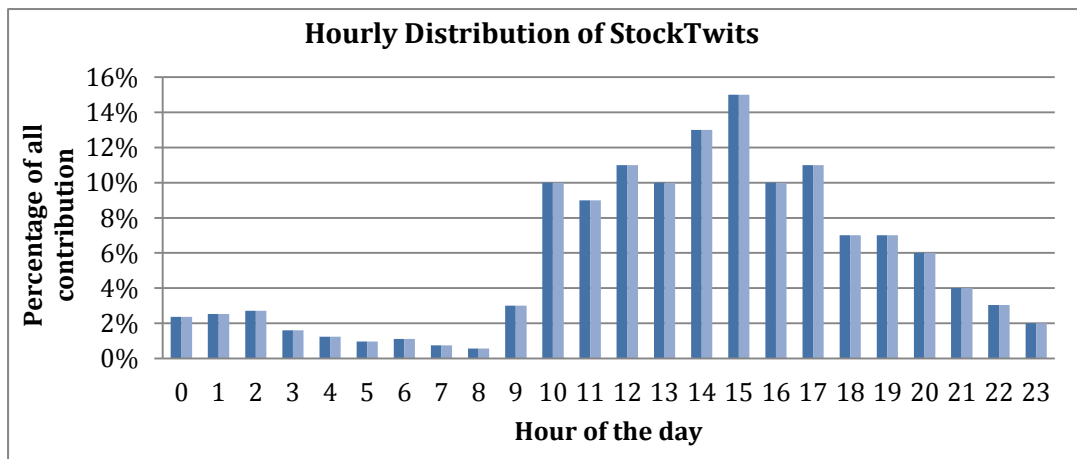


Figure 6.2: The hourly distribution of StockTwits

As it can be seen from Figure 6.2, message posting is concentrated between 10:00am and 5:00pm. This suggests a high activity by day traders during market times; hence, sentiments are most likely to develop more during market hours. The sentiments of day traders may have different influences at different times of the day

Chapter Six: Empirical Finance Analysis and Discussion

between the opening and closing times and the rest of the day. There are several possible explanations for this. For example, sentiments during the opening time of the market may be influenced by the actual trading activities and the real-time market fluctuations. Another possible explanation is that other market influences such as analysts' recommendations and other financial advisors who are likely to be seen during the market hours, may strongly affect investors' sentiments. On the other hand, sentiments after the market has closed are more likely to be based on investors' logical and intuitive analysis of the financial information available to them.

6.3.3 Distribution of StockTwits by Day of the Week

It is interesting at this point to investigate whether the volume of tweets might be distributed contrarily on different days of the week. Figure 6.3 shows the distribution of the tweet messages during the days of the week. Consistent with previous studies, it is observed that high volumes of tweets posted during working days reached a peak on Thursdays and showed low activity during the weekends and public holidays. This accords with Oh and Sheng's finding (2011) that the message posts reached a high level of activity during working days and a low volume during weekends.

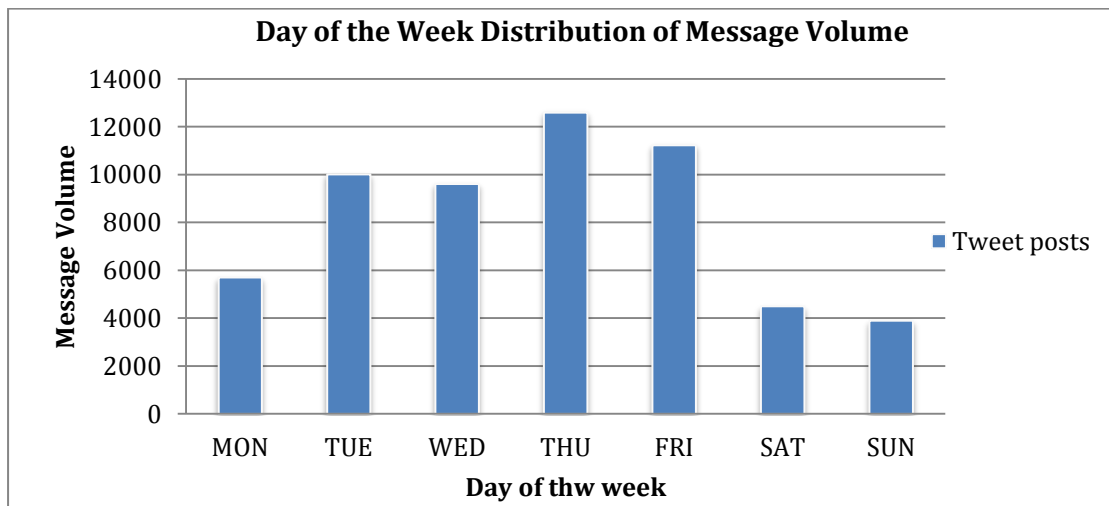


Figure 6.3: The distribution of StockTwits posts throughout the week

The foregoing discussions of findings provide evidence to support the assertion of this research that the diffusion and distribution of StockTwits postings is highly expressive of investors' behaviour in the stock market.

6.3.4 Distribution of StockTwits postings over the sample period

A graphical inspection in Figure 6.4 suggests that the StockTwits postings for the DJIA firms are reasonably stable over the considered period of April 2012 to April 2013. Nonetheless, some increase in the volume of posting activity is observed during the early summer, autumn months (i.e. Halloween), Christmas and New Year's Eve, suggesting that people tend to post more actively during these special occasions.

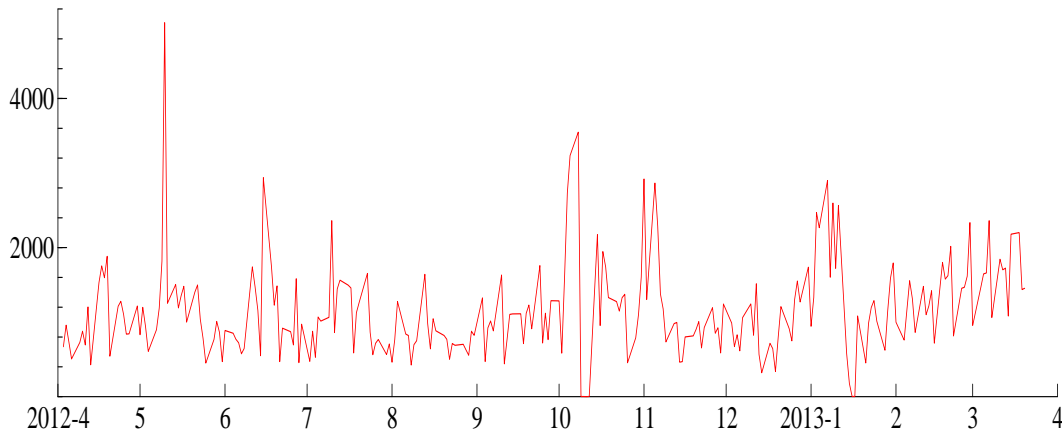


Figure 6.4: The daily StockTwits messages. Posting activity of 30 companies of DJIA index combined.

6.4 The Contemporaneous Relationship between Stock Micro-blogging Features and Stock Market Indicators

6.4.1 Pairwise Correlation

The initial significant effect of the relationships between stock micro-blogging features and the stock market is illustrated by the simple pairwise correlation matrix in Table 6.5. The Pearson correlation coefficients provide the statistically significant differences from zero between pairs of variables. Some of the reported results of correlations have been addressed in the early literature. For example, the relationships between different stock market variables such as trading volume and volatility are heavily documented correlations, and this is very significant in this study's result of 0.059. This correlation is very similar to that found in the seminal work by Antweiler and Frank (2004b) (0.063) in their study of stock message boards. The correlation between trading volume and number of messages in the sample shows a slightly weaker relationship of 0.294 compared to 0.322 on stock message boards. A very significant negative correlation is found between trading volume and the level of

Chapter Six: Empirical Finance Analysis and Discussion

agreement of StockTwits messages, which is a correlation that has not been found on the stock message boards. In general, the results of the pairwise correlations matrix suggest that there are statistically significant relationships between stock micro-blogging features and stock market measures and these relationships are worth examining further.

Table 6.5: Pearson correlation matrix

	Return	Trading Volume	Volatility	Bullishness	Message Volume	Agreement
Return	1.0000 -----					
Trading Volume	-0.0307*** (0.0077)	1.0000 -----				
Volatility	-0.0450*** (0.0001)	0.0588*** (0.0000)	1.0000 -----			
Bullishness	0.0027 (0.8174)	-0.1097*** (0.0000)	0.0355*** (0.0020)	1.0000 -----		
Message Volume	-0.0098 (0.3962)	0.2938*** (0.0000)	0.0234** (0.0422)	0.1441*** (0.0000)	1.0000 -----	
Agreement	-0.0032 (0.7835)	-0.0481*** (0.0000)	-0.0044 (0.7008)	0.0485*** (0.0000)	-0.1016*** (0.0000)	1.0000 -----

Note: This table reports the Pearson correlation of all the variables. The probabilities are shown in parentheses. *** and ** indicate significance at the 1% and 5% levels, respectively.

The pairwise correlations implied interesting relationships between StockTwits features and market indicators. A very strong correlation is observed between logged message volumes and logged trading volume ($r= 0.294$, $p= 0.000$). This exhibited correlation is the strongest among all correlations, indicating highly intensive message postings at the time when trading was conducted, as people tend to post more messages during the market hours of the actual trading activities. This provides support for the hypothesis (H1a) that people tend to discuss stocks that are traded more heavily. In contrast with earlier research studies (e.g., Wysocki, 1998), this research found no evidence of an association between message volume and stock returns. Although a negative relationship is found between message volume and stock market return, the result is not statistically significant ($r= -0.010$, $p=0.396$) (H1b).

Chapter Six: Empirical Finance Analysis and Discussion

This result is in line with Antweiler and Frank (2004b), who found a negative but insignificant correlation between the Raging Bull message boards and returns. A statistically significant and positive correlation is found between message volume and volatility ($r= 0.0234$, $p=0.000$), supporting (H1c). This implies that people tend to post more messages in the period of high market volatility, which may suggest that uncertainty causes investors to investigate the reasons behind this volatility with financial analysts and their relative peers. Our result is in line with Sprenger et al. (2014), who observe an increase in volatility as message volume rises.

The results show a statistically significant but negative correlation between bullishness and trading volume ($r= -0.110$, $p=0.000$) (H2a). To examine the hypothesis H2b, an interesting question needs to be asked: Can bullishness really predict returns? As it is well known, stock market returns are difficult to predict. In the simple pairwise correlations, a positive association is found between bullishness and stock return; however, this association is not statistically significant ($r= 0.003$, $p=0.856$) (H2b). On the other hand, a positive and significant relationship is found between bullishness and stock return volatility ($r= 0.036$, $p=0.000$) (H2c). Since the pairwise correlations measure the contemporaneous (short-term association) relationships between pairs of variables, it is worth noting that these correlations may be significant in the long run. This long-run association between tweet features and stock market measures will be tested using the Granger Causality test where the lagged relationships will be explored and discussed, as well as the direction of the effect of these relationships.

This study perceived a relatively strong negative correlation between agreement and trading volume ($r= -0.048$, $p=0.000$) (H3a). This suggests that stronger agreement among messages in a given period is associated with fewer trades during that period. This finding is also in line with Antweiler and Frank (2004b) and Sprenger et al. (2014), who found that the trading volume decreases as agreement increases. Is agreement associated with returns? As with bullishness and message volume, agreement also fails to explain stock returns; despite their negative relationships ($r= -0.003$, $p=0.784$) (H3b), agreement is not statistically significant. In line with Sprenger et al. (2014), this study fails to accept the hypothesis that agreement can explain stock return volatility ($r= -0.004$, $p= 0.701$) (H3c).

Chapter Six: Empirical Finance Analysis and Discussion

To conclude, the contemporaneous relationships between StockTwits features and stock market indicators provided in the pairwise correlations matrix are large and existing. The correlations between message volume and trading volume and between bullishness and trading volume appear to be the most significant and robust. The magnitude of these relationships suggests that they may yield information that is worth examining in a more sophisticated manner, applying rigorous econometric models to explore further unviable relationships that may exist.

6.4.2 Contemporaneous Regression

While the contemporaneous relations between tweet features and market variables provided in Table 6.5 have proposed exciting relationships, it is worth exploring these relationships in depth by addressing their independence.

Given the structure of the data, contemporaneous regressions in Eqs. (6.1), (6.2) and (6.3) are estimated for each of the financial variables separately using panel data. There are certain advantages associated with the use of panel data. Firstly, this approach enhances both the quality and quantity of data and allows more accurate model to control the impact of omitted variables. The fact that this research covers a relatively short time series (one year) makes the use of panel data a more suitable approach here since it allows the study of the dynamics of the variables of interest in a short time span. Since the DJIA index is composed of 30 companies, the data have dynamic effects where the intercept of the equations may differ according to the firms, the inclusion of firm dummies diminishes the biases in the estimation, and the time dummies capture time effects of the data. Therefore, panel regressions with cross-sectional fixed effects for each company are estimated using standard ordinary least squares (OLS) techniques, where the financial variables are treated as dependent variables and are regressed on three independent variables from the StockTwits: the bullishness index (\bar{B}_{it}^{**}), the message volume (M_{it}) and the level of agreement (A_{it}). The market return of the INDU index (DJIA index return) is added to the regressions as a control variable to control for overall market-wide effects²⁴. The OLS regression equations for each of the financial variables are shown in Table 6.6 and can be expressed as:

²⁴ In all reported results of this paper, the difference of log of INDU (INDU return) is employed as a control variable in all of the regressions that explain stock market variables.

Chapter Six: Empirical Finance Analysis and Discussion

$$R_{it} = \alpha_i + \beta_1 \bar{B}_{it}^{**} + \beta_2 M_{it} + \beta_3 A_{it} + \beta_4 MKT_t + \varepsilon_{it} \quad (6.1)$$

$$TV_{it} = \alpha_i + \phi_1 TV_{it-1} + \beta_1 \bar{B}_{it}^{**} + \beta_2 M_{it} + \beta_3 A_{it} + \beta_4 MKT_t + \varepsilon_{it} \quad (6.2)$$

$$\Delta V_{it} = \alpha_i + \phi_1 \Delta V_{it-1} + \beta_1 \bar{B}_{it}^{**} + \beta_2 M_{it} + \beta_3 A_{it} + \beta_4 MKT_t + \varepsilon_{it} \quad (6.3)$$

where, R_{it} , TV_{it} and ΔV_{it} are the stock market variables indicating returns, trading volumes and volatility respectively²⁵, whereas \bar{B}_{it}^{**} , M_{it} and A_{it} are the StockTwits features denoting bullishness, message volumes and agreement accordingly. MKT is the stock market index (INDU index) added to the regression to control for the market-wide effect. The OLS estimates of the coefficients β_s in Eqs (6.1-6.3) are the primary focus of these contemporaneous regression equations. These coefficients describe the dependence of the financial variables on the StockTwits features. Table 6.6 summarises the estimates of β_s .

Table 6.6: Contemporaneous Regressions

On data measured on daily frequency, panel regressions with company fixed effects are estimated separately for each market feature return (R_t), trading volume (TV_t) and volatility (ΔV_t) used as a dependent variable. The independent variables were obtained from StockTwits: bullishness index, the log number of messages and the level of agreement indicated by the coefficient $\beta_1, \beta_2, \beta_3$ and β_4 respectively. This table shows the predictive power of StockTwits features in explaining financial market indicators. In all regressions, market return is added as a control variable. Market return denotes the log difference of INDU price. N= 7560 company trading days.

	α_i	β_1	β_2	β_3	β_4	R^2	Durbin-Watson
R_{it}	-0.0155 (0.0361)	0.0924 (0.1080)	-0.0022 (0.0105)	-0.0134 (0.0479)	0.9975*** (0.0144)	0.3923	1.9916
TV_{it}	10.2822 (0.1594)	0.1261*** (0.0423)	0.0451*** (0.0042)	-0.0619*** (0.0188)	-0.0155*** (0.0056)	0.8430	2.0395
ΔV_{it}	-0.3081 (0.0398)	0.6126*** (0.1187)	0.0523*** (0.0116)	-0.0793 (0.0527)	-0.0727*** (0.0158)	0.2218	2.2367

Notes: * p<0.1, ** p<0.05, *** p<0.01, standard errors are in parentheses below the coefficients.

The stock market is known to be difficult to predict due to the fact that stock market data are noisy and time varying in nature. The most common question that one would need to ask in this context of study is as follows: Can stock micro-blogging messages (so called StockTwits) predict returns? The regression results of the return

²⁵The trading volume and volatility regressions both suffers from serial correlations therefore one period lag of trading volume and volatility are added in the model equations of trading volume and volatility respectively to get off the serial correlation problem.

Chapter Six: Empirical Finance Analysis and Discussion

equation as shown in Table 6.6 are largely as one would expect in an informationally efficient market. The p-value, for the null hypothesis that the current values of (all) tweet features do not forecast returns, is large, which strongly implies that neither bullishness nor message volume nor agreement exerts any statistically and economically significant influence in predicting stock returns. Thus, the findings support for the non-significant relationships between stock return and other tweet measures that have been previously found and reported in the pairwise correlations in Table 6.5. The contemporaneous regression that explains the trading volume implies that tweet variables contain relevant information that is not yet reflected in the volume of trade.

The trading volume regression results indicate that message posting and agreement level both exert a statistically and economically significant influence in forecasting the contemporaneous trading volume ($\beta_2 = +0.0451$ and $\beta_3 = -0.0619$, p-value < 0.001 for message volume and agreement respectively). This means that if the volume of the messages increases on an average by one message, this would lead to a 4.51% increase in the amount of traded shares of that particular discussed stock. These findings strengthen the theory that a high volume of message postings is associated with more trades where, in a given period, people tend to discuss stocks that are traded more heavily in the stock market during that period. The impact of the agreement on trading volume is challenging to predict, yet these results are in line with the studies by Antweiler and Frank (2004b) and Sprenger et al. (2014), where a negative correlations is observed between agreement and trading volume. As indicated by the coefficient $\beta_3 = -0.0619$, when the standard deviation of the buy and sell messages decreases by one unite of standard deviation²⁶ (As discussed in Chapter 4 in Section 4.8.6 that the level of agreement is measured by the standard deviation between the sell and buy message), the trading volume would decrease by 6.19%. This suggests that greater agreement between bearish and bullish messages in a period is associated with fewer trades during that period.

The bullishness variable offers a greater ability to predict stock trading volume, and its estimated coefficient is found to be positively significant at the 1% level where the magnitude of this significance relation is relatively large, as indicated by a

²⁶ Note that an increase in the standard deviation is interpreted as a high disagreement. Therefore, a decrease in standard deviation is indicated that the agreement is high

Chapter Six: Empirical Finance Analysis and Discussion

coefficient value of +0.1261. This implies that if the bullishness level increased on an average by one bullish investor, there will be 12.61% increase in the trading volumes of that particular stock. This result, however, suggests the existence of “positive-feedback trades”, theorised by Lakonishok et al. (1992), which implies that an optimistic investor is more likely to trade more while a pessimistic investor may trade less. This finding therefore supports Aitken (1998), who found that more liquidity is provided in good times and less in bad times.

As for the volatility regression, the highly significant positive coefficient of the bullishness measure in the volatility regression ($\beta_1 = +0.6126$, p-value = <0.001) means that when bullishness level increase on an average by one bullish investor, the change in volatility would increase by 0.6126%. This implies that the more bullish the message of a stock, the greater the market volatility of that stock for the same given period of time. This might be interpreted from the fact that bullish investor always tends to overvalue stocks by bringing its price above fundamental values where these price deviations would create risk or volatility in capital markets. It is also found that high message volumes trigger an increase in stock return volatility. A possible explanation of these is that when new messages are posted about a particular stock in the forum, these provide an indication that new information has arrived into the market about that particular stock. This new information might be good or (bad) news in nature which will cause investors to overreact or (underreact) to news by excessively buying or (selling) stocks in the market. These excessive trading activities of noise trading may create risk that causes the price to deviate from its expected levels. This is in line with the noise trading theory of De Long et al. (1990) who argues that the interactions of the trading activities of noise traders and arbitrageurs in capital markets can create risk called “noise traders’ risk” that cause prices to deviate from their fundamental levels and create limits to arbitrageurs. However, the findings show that the agreement level provides no statistical evidence to explain stock return volatility in the contemporaneous relationship.

From the results of the contemporaneous regressions reported in Table 6.6, this thesis concludes that none of the StockTwits measures offer any statistically significant ability to predict stock market returns. While the trading volume and volatility regressions show more robust results with respect to the predictive ability of almost all StockTwits variables, (e.g. the bullishness index shows the strongest effect

Chapter Six: Empirical Finance Analysis and Discussion

in anticipating market variables such as: trading volume and volatility). Yet, for trading purposes the contemporaneous correlations contribute little to the understanding of these prospective relationships. However, anticipating the subsequent changes in these relations is much more critical than predicting the contemporaneous correlations. Therefore, the next section will explore further the ability of StockTwits features to predict subsequent changes in the market indicators (and vice versa).

6.5 The Lead-Lag Relationship between Stock Micro-blogging Features and Stock Market Indicators

The pairwise correlations, represented by the contemporaneous correlations between StockTwits measures and financial market indicators, are highly significant. These results, however, raise a point of discussion of whether market movement causes StockTwits features or StockTwits causes stock price movements in the capital market. The economic importance of the anticipated relationships is not fulfilled by statistical significance alone. Therefore, in order to establish the empirical context for the impact of new information (StockTwits posts) on the market prices, this study first needs to understand and evaluate the nature of the relationships and their effects by analysing all possible relationships in both directions. To verify this hypothesis, one needs to explore the lagged associations between the tweets and the stock market features by estimating Vector Auto Regressions (VARs) models and making use of Granger causality test analysis. This test does not imply causality; rather, it investigates the statistical pattern of lagged correlations by evaluating the bidirectional effect of StockTwits and the stock market. In the financial market, the ability to anticipate subsequent changes is more critical than contemporaneous associations. The following section focuses on the interrelation between tweets measures and stock market indicators using VAR analysis.

- **Estimated Vector Auto Regressions (VARs) Model**

To study in more detail the correlation between StockTwits and financial market, a time sequencing test is performed to examine the lead-lag relationships between StockTwits variables and financial market indicators. Hence, this section models the short-term interrelation by modelling VAR for each of the three financial

Chapter Six: Empirical Finance Analysis and Discussion

variables separately. Thereafter, Granger Causality tests (Granger, 1969) show whether StockTwits variables affect the financial market or vice versa. Information criteria, namely likelihood ratio (LR), Schwarz Bayesian Information Criterion (SBIC), Akaike (AIC), and Hannan-Quin Information Criterion (HQIC), will determine the appropriate lag structure in the VARs. Likelihood ratio and Akaike information criteria for the model order indicate that several additional lags are needed, while SBIC and HQIC prefer smaller model orders. Since three independent measures of StockTwits will be included in all VARs models, a highly complex model is likely to be estimated if many lags are included in the system²⁷. Therefore, the smaller model orders are used where SBIC and HQIC will be chosen to determine the appropriate number of lags as both favour a less complex model.

Three VARs models will be estimated, where the endogenous variables in each of the VARs are one of the market variables (Return, Trading Volume and Volatility) and the three tweet features (bullishness, message volume and agreement). The exogenous variables in each VAR include lags of their own and lags of other endogenous variables, dummy variables for first day of the trading week (NWK) to control for the potential return anomaly effect²⁸ in line with Antweiler and Frank (2004b), and dummy variables controlling for market-wide effect (INDU index). The following subsections will discuss the VARs system for return, trading volume and volatility respectively.

6.5.1 VAR - Return Model

In this section, the VAR model is estimated with the tweet measures and market returns. The ultimate purpose is to investigate how tweet features and stock market returns interact and identify the (statistical) causality between StockTwits features and returns.

$$Y_{it} = \alpha_i + \sum_{i=1}^p \beta_i Y_{t-i} + MKT_t + NWK_t + \varepsilon_{it} \quad (6.4)$$

²⁷Akaike Information Criteria (AIC) (Akaike, 1974) usually favours more complex models; therefore, other information criteria will be more appropriate to determine the lag length in our research study, hence making the model easy to interpret.

²⁸The day-after-holiday effect is one of the stock return anomalies where returns of stocks are found to be lower, e.g. on Mondays than for other days of the week (Thalar, 1987). To control for this return anomaly a dummy variable is created whereby this dummy takes a value of one on the day after a holiday and zero otherwise.

Chapter Six: Empirical Finance Analysis and Discussion

where, Y_{it} is a vector of the stock returns (R_{it}) and tweet features, specifically bullishness (\bar{B}_{it}^{**}), message volume (M_{it}) and agreement (A_{it}) for a company i at time t , respectively. α_i is a vector of intercepts. β_i is a (4 x 4) matrix of the estimated parameters, with the diagonal parameters capturing the autoregressive terms, while the off-diagonal ones capturing the Granger causality between the variables in the system. ε_i is a vector of innovations. MKT is the market return of INDU index (DJIA index return) added to the regressions as a control variable to control for the overall market-wide effects, and NWK is a dummy variable for the first day of the new trading week to control for potential return anomaly effect. For reversed causal relations, the tweet features are the dependent variables and return is the independent variable in each equation of tweet features. As mentioned earlier in the previous section, a simple VAR model is preferred to a complex system and the lag length is determined based on the SBIC and HQIC information criteria. As expected, the SBIC favours the inclusion of two lags, whereas HQIC requests three additional lags. Therefore, to get a more complete picture of the associated relationship, HQIC would be chosen in favour of SBIC as it favours a longer lag structure of five lags (i.e., $p=5$). Hence, a VAR with five lags is estimated, and Granger Causality tests are carried out.

Table 6.7 reports the outcomes from estimating VAR-returns for five lags. The blocks of rows indicate the contribution of each independent variable at lags 1, 2, 3, 4 and 5. To test the joint significance of the lagged values of a given independent variable, the p-values are obtained by estimating each equation in the VAR system separately using ordinary least square (OLS) techniques.²⁹ All estimated equations are based on panel regressions with company fixed effect.

As it can be seen from Table 6.7, none of the StockTwits measures has a lead and lagged effect in explaining stock market return. This however, supports the previous research arguments that stock market returns are difficult to predict and it is concluded that StockTwits measures do not contain any valuable information for predicting stock market return. Nevertheless, since VAR analysis allows to investigate the predictive power in two directions (from StockTwits to stock return and from stock return to StockTwits), it is therefore important to investigate whether stock return might have any predictive ability in explaining tweet features. The sophisticated VAR analysis reveals

²⁹ It is assumed that disturbance terms in the VARs system are independent by which the disturbance terms in any variable equations have no obvious relation to disturbances in other equations. Relaxing the assumptions of independence of error terms across equations does not affect the results.

Chapter Six: Empirical Finance Analysis and Discussion

that the bullishness index shows no significant associations with stock return in either direction. This is unfortunately bad news for market participants trying to use the bullishness measure extracted from the online StockTwits forum for their short-term market timing strategies.

Table 6.7: Result of the VAR and Granger causality tests for Stock Return

VAR-Return and StockTwits (Five lags)					
Dependent variable					
Independent variable	Lag	Return (R_t)	Bullishness (\bar{B}_{it}^{**})	Message Volume (M_t)	Agreement (A_t)
Return	R_{t-1}	-0.0036	-0.0005	0.0166*	-0.0053**
	R_{t-2}	-0.0140	-0.0007	-0.0187**	-0.0023
	R_{t-3}	0.0054	-0.0004	-0.0049	-0.0057***
	R_{t-4}	-0.0138	-0.0008	0.0169*	-0.0007
	R_{t-5}	-0.0137	0.0012	0.0260**	-0.0015
Bullishness	\bar{B}_{t-1}^{**}	-0.1708	0.3004***	0.4137***	0.0650**
	\bar{B}_{t-2}^{**}	0.1219	0.0718***	-0.1173	0.0545*
	\bar{B}_{t-3}^{**}	0.0625	0.0337***	-0.1056	-0.0037
	\bar{B}_{t-4}^{**}	0.0179	0.0390***	0.2115*	0.0437
	\bar{B}_{t-5}^{**}	0.0681	0.0388***	0.2901**	0.0218
Message Volume	M_{t-1}	-0.0004	0.0032***	0.3567***	-0.0063**
	M_{t-2}	0.0047	0.0011	0.1114***	-0.0013
	M_{t-3}	-0.0004	0.0007	0.0072	-0.0052*
	M_{t-4}	-0.0063	-0.0011	-0.0319**	-0.0022
	M_{t-5}	-0.0059	-0.0004	0.0112	0.0007
Agreement	A_{t-1}	-0.0450	0.0024	-0.0435	0.0681***
	A_{t-2}	0.0240	-0.0005	-0.0309	0.0266**
	A_{t-3}	-0.0370	-0.0039	-0.0125	0.0140
	A_{t-4}	-0.0573	0.0036	-0.0132	0.0263**
	A_{t-5}	0.0443	-0.0006	-0.0578	0.0106
Constant		0.0048	0.1543***	1.2881***	0.0736***
Market (MKT_t)		0.9939***	-0.0022	-0.0402***	0.0002
Dummy (NWK_t)		0.0072***	-0.0174***	-0.4047***	0.0248***
R-squared (R^2)		0.3872	0.6291	0.5433	0.0380
N-Observation		7,410	7,410	7,410	7,410
Granger causality test (χ^2)		10.4881 (0.7880)	56.3097*** (0.000)	41.4414*** (0.0003)	59.5130*** (0.000)

Unlike Wysocki (1998) and Antweiler and Frank (2004b), who found a predictive ability in the number of messages on the stock message board in predicting stock returns,

Chapter Six: Empirical Finance Analysis and Discussion

this study has failed to uncover any explanatory power in this direction. Our results are in line with a similar study by Sprenger et al. (2014) who do not find message volume to be related to stock return and suggest that investors may take a more nuanced approach in processing information content of stock micro-blogs compared to message boards. However, the inverse causality is strongly evident in the results of this research. Stock returns of one and two days prior seem to predict current-day message volume. The results, as shown in Table 6.7, indicate that a significant positive relationship between the return on day t and the number of tweet messages on day $(t+1)$ is statistically present ($c=+0.017$, $p\text{-value}< 0.10$). On the subsequent day $(t+2)$, the effect reverses itself and a negative coefficient of a similar magnitude effect is likely to be found ($c=-0.019$, $p\text{-value}< 0.05$). These two tiny estimated effects are almost identical in value, have opposite signs and are offsetting (likely to cancel each other out). The partial effect ($+0.017$) implies that a 100% increase in stock prices leads to an almost 2% increase in message posting. This result might be explained by the fact that increases in stock returns occur when news of a good nature arrives in the market, causing people to post more messages to diffuse and discuss such good information and exchange ideas with their peers in the forum. Another possible explanation is that some of the companies and other financial analysts may regard StockTwits as a forum to disseminate their financial information and news, such as earnings announcements and financial performance reports, which causes people to discuss such events with their brokers and peers.

In line with Antweiler and Frank (2004b) and Sprenger et al. (2014), agreement has no explanatory power to explain stock return. The opposite relationship, however, does hold. It is found that there is a negative coefficient on the return on day $(t+1)$ and day $(t+3)$. This implies that higher returns lead to disagreement among traders ($c= - 0.0053$, $p\text{-value}< 0.05$ and $c=-0.0056$, $p\text{-value}<0.001$ for the first and third lags respectively). This result is inconsistent with Berkman et al. (2009) and Irvine and Giannini (2012) who find that divergence of opinion is significantly negative relative to returns. One possible explanation for this result is as follows: When an asset experiences a high return, there may be present two types of speculators, namely momentum traders, who follow a momentum strategy (buy high/sell low) and are very optimistic about buying an asset at a higher price, and contrarian traders, who follow a contrarian strategy (buy low/sell high) and are willing to sell that asset at a higher price; these traders' mutual trading in the

Chapter Six: Empirical Finance Analysis and Discussion

market will cause differences of opinion about the value of that traded asset among each type of trader.

Taken together, the result of VAR-return reveals that there is no evidence to suggest that tweet features might predict subsequent market return. Neither bullishness nor the number of messages nor the extent of disagreement today forecasts tomorrow's return. The Granger causality tests fail to reject the null hypothesis of no predictability in stock returns for all StockTwits measures. However, causality is still running from the reverse direction for both the message volume and level of agreement (except bullishness) where the effect from stock returns to StockTwits features is in a much more significant direction, indicated by the high magnitude of Chi-square tests.

6.5.2 VAR - Trading Volume Model

Trading volume is one of the most important measures of the financial market and it has received attention in most of the empirical finance literature. As with stock return, a VAR model with StockTwits variables and trading volume is estimated and Granger causality tests are carried out. The VAR equation below describes this relationship:

$$Y_{it} = \alpha_i + \sum_{i=1}^p \beta_i Y_{t-i} + MKT_t + NWK_t + \varepsilon_{it} \quad (6.5)$$

where, Y_{it} is a vector of the stock trading volume (TV_{it}) and tweet features, specifically bullishness (\bar{B}_{it}^{**}), message volume (M_{it}) and agreement (A_{it}) for a company i at time t , respectively. α_i is a vector of intercepts. β_i is a (4 x 4) matrix of the estimated parameters, with the diagonal parameters capturing the autoregressive terms, while the off-diagonal ones capturing the Granger causality between the variables in the system. ε_i is a vector of innovations. MKT is the market return of INDU index (DJIA index return) added to the regressions as a control variable to control for the overall market-wide effects, and NWK is a dummy variable for the first day of the new trading week to control for potential return anomaly effect. Using SBIC information criteria, a VAR system with six lags (i.e., $p=6$) is estimated. The results from estimating VAR-trading volume are shown in Table 6.8.

As it can be seen from Table 6.8, the first block of rows shows that trading volume is a powerful predictor of itself. All lags (except for lags 4 and 6) are positive and significant at the 5% level. These results are not surprising, however, because

Chapter Six: Empirical Finance Analysis and Discussion

what is interesting in the data presented in this table is the question of whether or not tweet features can predict trading volume and vice versa.

Table 6.8: Result of the VAR and Granger causality tests for Trading Volume

VAR-trading volume and StockTwits (Six lags)					
Independent variable	Lag	Dependent variable			
		Trading Volume (TV _t)	Bullishness (\bar{B}_t^{**})	Message Volume (M _t)	Agreement (A _t)
Trading Volume	TV _{t-1}	0.2972***	0.0106***	0.3208***	-0.0040
	TV _{t-2}	0.0618***	-0.0074**	-0.1344***	-0.0124*
	TV _{t-3}	0.0329***	-0.0002	0.0267	0.0046
	TV _{t-4}	-0.0150	-0.0063*	-0.0504	0.0123
	TV _{t-5}	0.1055***	0.0014	-0.0043	-0.0028
	TV _{t-6}	0.0246**	-0.0028	-0.1290***	-0.0045
Bullishness	\bar{B}_{t-1}^{**}	-0.0653	0.2962***	0.3461***	0.0661**
	\bar{B}_{t-2}^{**}	0.0224	0.0731***	-0.0823	0.0589**
	\bar{B}_{t-3}^{**}	0.0174	0.0331***	-0.1289	-0.0036
	\bar{B}_{t-4}^{**}	0.0008	0.0398***	0.2124*	0.0456
	\bar{B}_{t-5}^{**}	0.1091**	0.0371***	0.2678**	0.0273
	\bar{B}_{t-6}^{**}	-0.0963**	0.0030	0.0708	-0.0171
Message Volume	M _{t-1}	0.0038	0.0029**	0.3418***	-0.0056**
	M _{t-2}	-0.0033	0.0014	0.1180***	-0.0009
	M _{t-3}	-0.0020	0.0009	0.0108	-0.0063**
	M _{t-4}	-0.0276***	-0.0008	-0.0291**	-0.0032
	M _{t-5}	0.0105**	-0.0005	0.0219*	0.0014
	M _{t-6}	0.0147***	-0.0001	-0.0074	-0.0005
Agreement	A _{t-1}	-0.0337*	0.0034	-0.0264	0.0696***
	A _{t-2}	-0.0561***	-0.0005	-0.0227	0.0256**
	A _{t-3}	-0.0149	-0.0036	-0.0005	0.0141
	A _{t-4}	0.0045	0.0029	-0.0181	0.0273**
	A _{t-5}	0.0208	-0.0009	-0.0625	0.0086
	A _{t-6}	0.0369**	0.0011	0.0113	0.0047
Constant		7.2498***	0.2228***	0.8656	0.1748
	Market (MKT_t)	-0.0217***	-0.0022	-0.0461***	0.0010
	Dummy (NWK_t)	-0.1437***	-0.0200***	-0.4412***	0.0268***
R-squared (R²)		0.8479	0.6293	0.5494	0.0377
N-Observation		7,380	7,380	7,380	7,380
Granger causality test (χ²)		91.1161*** (0.000)	65.3595*** (0.000)	167.7794*** (0.000)	57.7570*** (0.000)

Chapter Six: Empirical Finance Analysis and Discussion

Table 6.8 shows that bullishness predicts stock trading volume. The bullishness sentiment exerts a statistically and economically significant positive influence on the fifth day's trading volume ($c = +0.11$, $p\text{-value} < 0.05$). This finding is consistent with Baker and Stein (2004) and Liu (2006), who show that investor sentiments are positively correlated with stock market liquidity. On the subsequent day ($t+6$), however, the effect reverses itself where bullishness shows a negative predictability of trading volume but with a smaller magnitude effect ($c = -0.10$, $p\text{-value} < 0.05$). The interpretation of this reversal effect is twofold. First, the statistically significant positive impact of bullishness on the fifth day's trading volume might be interpreted from a theoretical viewpoint and relative norms of the positive-feedback trading behaviour (Lakonishok et al., 1992); i.e. bullish investors who are optimistic about the future development of the stock market may trade more on average than bearish investors with pessimistic beliefs (Aitken, 1998). This result implies that optimistic investors provided more liquidity in the stock market by triggering higher trading volumes. On the other hand, the negative relationship between bullishness and trading volume on day ($t+6$) suggests that bullish investors with high levels of optimism about stocks will tend to overvalue stocks by driving up prices from their fundamental level, which consequently lowers the subsequent returns (Brown and Cliff, 2004) and may result in reducing investors' desire to trade those particular stocks³⁰. The delay in the lead-lag effect between trading volume and bullishness may have resulted from the delayed stock price reactions to information arriving in the market. The reverse direction, however, confirms a very significant effect of trading volume in explaining bullishness, indicated by a positive coefficient of trading volume on day ($t+1$) ($c = +0.01$ $p\text{-value} < 0.001$). This implies that the more shares traded today, the more bullish tomorrow's messages. This result shows that past trading volumes have positive impacts on current bullishness; therefore, a high volume of trade triggers an increase in investor bullishness while low trading volume prompts a reduction in investor bullishness. Although an inverse effect appeared on day ($t+2$), this effect is very small in magnitude, as indicated by the small negative coefficient ($c = -0.007$).

Lag message volume shows a significant effect in explaining current-day trading volumes. This significance is driven by the fourth to the sixth lags of message

³⁰ Note that the direct relationship between bullishness and trading volume is hard to observe. However, it is observed indirectly through the effect of bullishness on stock return.

Chapter Six: Empirical Finance Analysis and Discussion

volume. Nevertheless, the reliability of this significant relationship is called into question by the absence of an effect from the first three lags. The negative (positive) effect of message volume on trading volume in forth (fifth and sixth) lags, may be interpreted as more messages posted the more sentiments investors are likely developed which resulted in increase trading. These means that increase and decrease in message posted would be interpreted as bullish (bearish) attitude toward a particular discussed stock in the forum. According to DSSW (1990), as investors become more bullish (bearish) about a particular asset, their demand to purchase (sell) that asset increased whereby increasing the volume of trade of that asset in capital market. At the same time, trading volumes seem to be a powerful predictor of the number of messages posted on StockTwits, denoted by the significant coefficient at lags one, two and six at the 1% level of significance. These findings are consistent with those of Sprenger et al. (2014), who found a bidirectional effect of causality between message volumes and trading volumes. The positive coefficient ($c = +0.3208$, $p\text{-value} < 0.01$) of trading volume on day ($t+1$) implies the tendency of investors to discuss stocks that are traded more heavily in the stock market. This result is particularly true for noise traders trying to manipulate the market and taking advantage of such manipulation strategies by bringing their own traded shares into discussions. Another possible explanation for this result is that small investors are more likely to gossip with and consult their peers about shares they wish to purchase or about purchases they have just made.

One of the most challenging relationships to anticipate, among all StockTwits/market relations, is the role of agreement in explaining trading volume in the stock market. This study produced results that corroborate the findings of a great deal of the previous work in this field. The findings support the traditional hypothesis of Harris and Raviv (1993) that disagreement induces trading, as indicated by the negative coefficients on the agreement index one and two days ago in the trading volume regression. This result implies that higher disagreement among traders will cause the market price of the stock to be relatively higher than its intrinsic value, which causes investors with optimistic beliefs to express their over-confidence through higher trading. In line with the contemporaneous regressions, greater agreement on day t is associated with fewer trades in the next few days. This signifies the explanatory power of agreement in explaining the trading volume (more

Chapter Six: Empirical Finance Analysis and Discussion

elaborated details about this perspective relationship will be provided in Chapter 7). The opposite relationship, however, does exist. It is worth remarking that the findings of this study provide evidence that trading volume also causes disagreement although this effect is quite limited (only significant at lag two) and economically small in magnitude ($c = -0.0124$, $p\text{-value} < 0.1$).

Looking down the first column, the impact of StockTwits features on trading volume is evident, giving the significant Chi-square statistic of 91.116. The Granger Causality test implies that tweet features appear to contain predictive information with respect to trading volume. In the data measured at daily frequency, the effect of StockTwits features (e.g. bullishness and agreement) on trading volume is more significant than in the reverse direction. Conversely, the Chi-sq. χ^2 values illustrate that the strongest effect is found in the direction from trading volume to message volume rather than in the reverse direction

6.5.3 VAR - Volatility Model

Stock micro-blogging is a platform where the sentiments of irrational investors and noise traders play an active role in information diffusion (Oh and Sheng, 2011). If this study assumes that noise traders engage in online conversations by actively posting messages, then their actions may induce market volatility. Furthermore, the pairwise correlations in Table 6.5 show that stock return volatility and trading volume are correlated. The previous section also showed that trading volumes and message volume are among the strongest correlations found between tweet measures and stock market indicator variables. Therefore, the aim of this section is to consider whether StockTwits measures provide any explanatory power in forecasting stock return volatility. As with return and trading volume, a VAR model with tweet measures and stock return volatility is estimated as follows:

$$Y_{it} = \alpha_i + \sum_{i=1}^p \beta_i Y_{t-i} + MKT_t + NWK_t + \varepsilon_{it} \quad (6.6)$$

where, Y_{it} is a vector of the change in volatility of stock returns (ΔV_{it}) and tweet features, specifically bullishness (\bar{B}_{it}^{**}), message volume (M_{it}) and agreement (A_{it}) for a company i at time t , respectively. α_i is a vector of intercepts. β_i is a (4 x 4) matrix of the estimated parameters, with the diagonal parameters capturing the

Chapter Six: Empirical Finance Analysis and Discussion

autoregressive terms, while the off-diagonal ones capturing the Granger causality between the variables in the system. ε_t is a vector of innovations. MKT is the market return of INDU index (DJIA index return) added to the regressions as a control variable to control for the overall market-wide effects, and NWK is a dummy variable for the first day of the new trading week to control for potential return anomaly effect. Using SBIC information criteria, VAR systems with six lags (i.e., $p=6$) are estimated and Granger causality tests are conducted. The results from estimating VAR-Volatility are shown in Table 6.9.

The results in Table 6.9 unsurprisingly show that stock return volatility is a powerful predictor of itself. Statistically significant negative coefficients of all lags are found at the 1% level. The data from this table reveal that bullishness has a significant influence in explaining market volatility, indicated by the strong negative coefficient of $c = -0.236$ at the 5% level of significance. This implies that the more bullish the investor messages today, the less volatile tomorrow's market will be. This suggests that there is a negative volatility spillover from the bullishness index of StockTwits to the market. This result concurs with Lee et al. (2002), who found a negative impact of the shift in sentiment on stock return volatility, whereby the bullish (bearish) shift in sentiments results in downward (upward) revision in volatility.

On the other hand, the opposite causal direction does exist, where the greater the market volatility today, the more bullish tomorrow's messages. Thus, there is a significant positive effect flowing from past market volatility to bullishness. The result of this study is in line with the findings of Antweiler and Frank (2004b), who show that a more bullish message is more likely to be found during a volatile period. One possible explanation for this significant positive relationship is the varying risk preference of investors where, in a period of high volatility risk-loving investors increase their demand for risky assets as they become more bullish by amplifying the level of market risk and thereby earning higher expected returns. Conversely, risk-averse investors will avoid trading in periods of high volatility as they become more bearish by reducing their level of market risk, thus resulting in lower expected returns. The results provide support for Baek et al. (2005), who demonstrate that a shift in risk attitudes of investors may explain short-term movement in asset prices better than any other fundamental factors. Our findings also match those of Yu and Yuan (2011), who

Chapter Six: Empirical Finance Analysis and Discussion

find that investor sentiments affect risk-return trade-offs whereby low (high) mean-variance trade-offs are observed in high (low sentiments) periods respectively.

Volume of posting two days ago appears to be strongly negatively correlated with market volatility. This implies that more message posting signifies significantly less market volatility ($c = -0.0268$, $p\text{-value} < 0.10$). This result indicates that the discussions and conversations taking place in the StockTwits forum via increased postings are more likely to reduce the level of uncertainty regarding the traded security in the capital market. On the other side, it is observed that high volatility triggers an increase in message volume, indicated by the significant coefficient of volatility at lags one, two, four and six in the volume regression of the VAR system. The result here concurs with Sprenger et al. (2014), who find that higher volatility leads to increased message posting. This an interesting result; however, while market volatility serves as a proxy for uncertainty, increasing message volume in a highly volatile period confirms that uncertainty causes investors to engage in conversations, exchanging opinions, asking questions and consulting their peers. Another possible explanation is that higher market volatility will result in extensive debate among market participants, which in turn results in releases of new information via the posting of more messages in the StockTwits forum. The VAR-volatility table also allows us to examine the strength of effects in each direction. The Granger block exogeneity tests indicate that market volatility has a more significant effect on message posting than in the reverse direction.

In contrast to earlier findings by Antweiler and Frank (2004b) and Sprenger et al. (2014), this study provides some confirmatory evidence that disagreement among traders is associated with higher market volatility ($c = -0.0834$, $p\text{-value} < 0.1$). This finding suggests that volatility may be a reflection of the divergence of opinions among market participants. This result may also be explained by the fact that the unpredictable fluctuations of noise trader sentiments and/or opinions about selling and buying a security may create risk, causing prices to diverge from fundamental values (market volatility). The activities of noise traders will cause prices to fluctuate above or below equilibrium, which will subsequently cause arbitrageurs to engage in trading by pushing prices back or forth to equilibrium to keep the market efficient. The present findings seem to be consistent with other research (DSSW, 1990; Campbell

Chapter Six: Empirical Finance Analysis and Discussion

and Kyle, 1993; Koski et al., 2004), which asserted that noise trading increases volatility and creates risk termed “noise trader risk”.

Table 6.9: Result of the VAR and Granger causality tests for Stock Return Volatility

VAR-volatility and StockTwits (six lags)					
Independent variable	Lag	Volatility (ΔV_t)	Bullishness (\bar{B}_t^{**})	Message Volume (M_t)	Agreement (A_t)
Volatility	V_{t-1}	-0.7502***	0.0026**	0.0897***	-0.0062**
	V_{t-2}	-0.5843***	-0.0003	0.0355**	-0.0070**
	V_{t-3}	-0.4922***	-0.0002	0.0238	-0.0017
	V_{t-4}	-0.3792***	-0.0004	0.0405**	-0.0025
	V_{t-5}	-0.2274***	-0.0004	0.0432***	-0.0042
	V_{t-6}	-0.1064***	0.0002	0.0044	0.0004
Bullishness	\bar{B}_{t-1}^{**}	-0.2363**	0.2961***	0.3256***	0.0714***
	\bar{B}_{t-2}^{**}	-0.0572	0.0746***	-0.0498	0.0555*
	\bar{B}_{t-3}^{**}	0.1220	0.0329**	-0.1126	-0.0069
	\bar{B}_{t-4}^{**}	-0.0398	0.0386***	0.1730	0.0510*
	\bar{B}_{t-5}^{**}	-0.0096	0.0382***	0.2777**	0.0270
	\bar{B}_{t-6}^{**}	-0.0344	0.0028	0.1113	-0.0217
Message Volume	M_{t-1}	0.0000	0.0031**	0.3481***	-0.0050**
	M_{t-2}	-0.0268**	0.0014	0.1197***	-0.0020
	M_{t-3}	-0.0178	0.0006	0.0083	-0.0064**
	M_{t-4}	-0.0134	-0.0010	-0.0360***	-0.0021
	M_{t-5}	-0.0176	-0.0006	0.0156	0.0009
	M_{t-6}	-0.0079	0.0000	-0.0028	-0.0010
Agreement	A_{t-1}	-0.0834*	0.0030	-0.0402	0.0691***
	A_{t-2}	0.0117	-0.0006	-0.0296	0.0257**
	A_{t-3}	0.0340	-0.0042	-0.0160	0.0147
	A_{t-4}	-0.0276	0.0033	-0.0160	0.0280**
	A_{t-5}	0.1110**	-0.0005	-0.0607	0.0085
	A_{t-6}	0.0439	0.0015	0.0125	0.0048
Constant		0.3027***	0.1546***	1.2815***	0.0766***
Market (MKT_t)		-0.0481***	-0.0020	-0.0441***	0.0003
Dummy (NWK_t)		-0.0955***	-0.0177***	-0.3930***	0.0241***
R-squared (R²)		0.3697	0.6289	0.5461	0.0380
N-observation		7,380	7,380	7,380	7,380
Granger causality test (χ^2)		32.3476** (0.0200)	57.8052*** (0.000)	92.4533*** (0.000)	60.2807*** (0.000)

Chapter Six: Empirical Finance Analysis and Discussion

Similarly, the findings of this study exhibited a significant negative coefficient of volatility in explaining the next-day agreement among messages. This implies that high market volatility induces disagreement among traders. There are several possible explanations for this result. For instance, fluctuations in an asset's price in the market may cause a high divergence of opinion between noise traders and arbitrageurs, who normally trade against each other. Furthermore, as suggested in the DSSW model of noise traders, changes in the noise traders' misperceptions of asset risk cause noise traders to follow one another in selling (buying) risky assets just when other noise traders are selling (buying). An increase in misperception of the asset's risk will cause a raise in price uncertainty that deters risk-averse arbitrageurs from holding that risky asset. The Granger causality tests indicate that, while there is a unidirectional effect running between agreement and volatility, the more significant flow is from the market volatility to agreement.

In summary, unlike the contemporaneous regression of volatility, which provides weak evidence that StockTwits features have an explanatory power in explaining stock return volatility, the results of VAR-volatility reveal that StockTwits features can anticipate the subsequent changes of volatility. Yet the effect of volatility on StockTwits is stronger than the effect of StockTwits on volatility.

6.6 Impulse Response Functions

Impulse response functions (IRF) refer to the analysis of the dynamic reaction of a system in response to any external changes, called impulses. They are a graphical representation output of a VAR model that helps to provide more insight into the dynamic relationship between the study variables in the system.

Given the VAR models' results (for returns, trading volume and volatility) in the previous section, this section analyses in more detail the dynamic linkages between stock market indicators (returns, trading volume and volatility) and StockTwits variables (bullishness, agreement and message volume). The Generalised Impulse Response Functions (GIRFs) of Pesaran and Shin (1998) are estimated for the cases where Granger causality is not rejected. The GIRFs are displayed in Figure 6.5 (a, b, and c) for VAR return, trading volume and volatility model, respectively. Overall, the results of the GIRFs (10 periods) from one standard error shock of the

Chapter Six: Empirical Finance Analysis and Discussion

variables in question are in line with the findings for Granger causality (See Tables 6.7, 6.8 and 6.9).

In the VAR return model (see Figure 6.5a), stock returns neither affect any of the StockTwits variables nor do they respond to any shocks in StockTwits variables. This is consistent with the VAR return system and confirms the unpredictability of stock returns using StockTwits data.

With regard to the VAR trading volume model (See Figure 6.5b), a one-standard-error shock to trading volume has a positive/negative impact on bullishness and message volumes on the first and second days. These positive/negative impacts are also found in the response of trading volume to the shock delivered by bullishness, being positive on the fifth and negative on the sixth day, and message volume, being negative on the fourth day and positive on the fifth and sixth days; this suggests that positive and negative impacts flow from either direction. On the other hand, a shock to trading volume has a negative impact on agreement on the first and second days, whilst a significant negative response of trading volume to a shock in agreement is found on the second day. This is consistent with the VAR-Trading volume model shown in Table 6.8.

In the volatility model (See Figure 6.5c), a one-standard-deviation shock to volatility results in an increase in bullishness (on the first day only) and message volume (on the first, second, fourth and fifth days), whilst negative correlations are found in the response of volatility to the shocks of bullishness (on the first day only) and message volume (significant on the second day only). It is also found that a one-standard-error shock to agreement leads to a depreciation of the stock return volatility on the first and second days. On the other hand, agreement responds negatively (positively) to a shock to stock return volatility on the first (fifth) day respectively, in line with the corresponding VAR model results. (See Table 6.9).

Chapter Six: Empirical Finance Analysis and Discussion

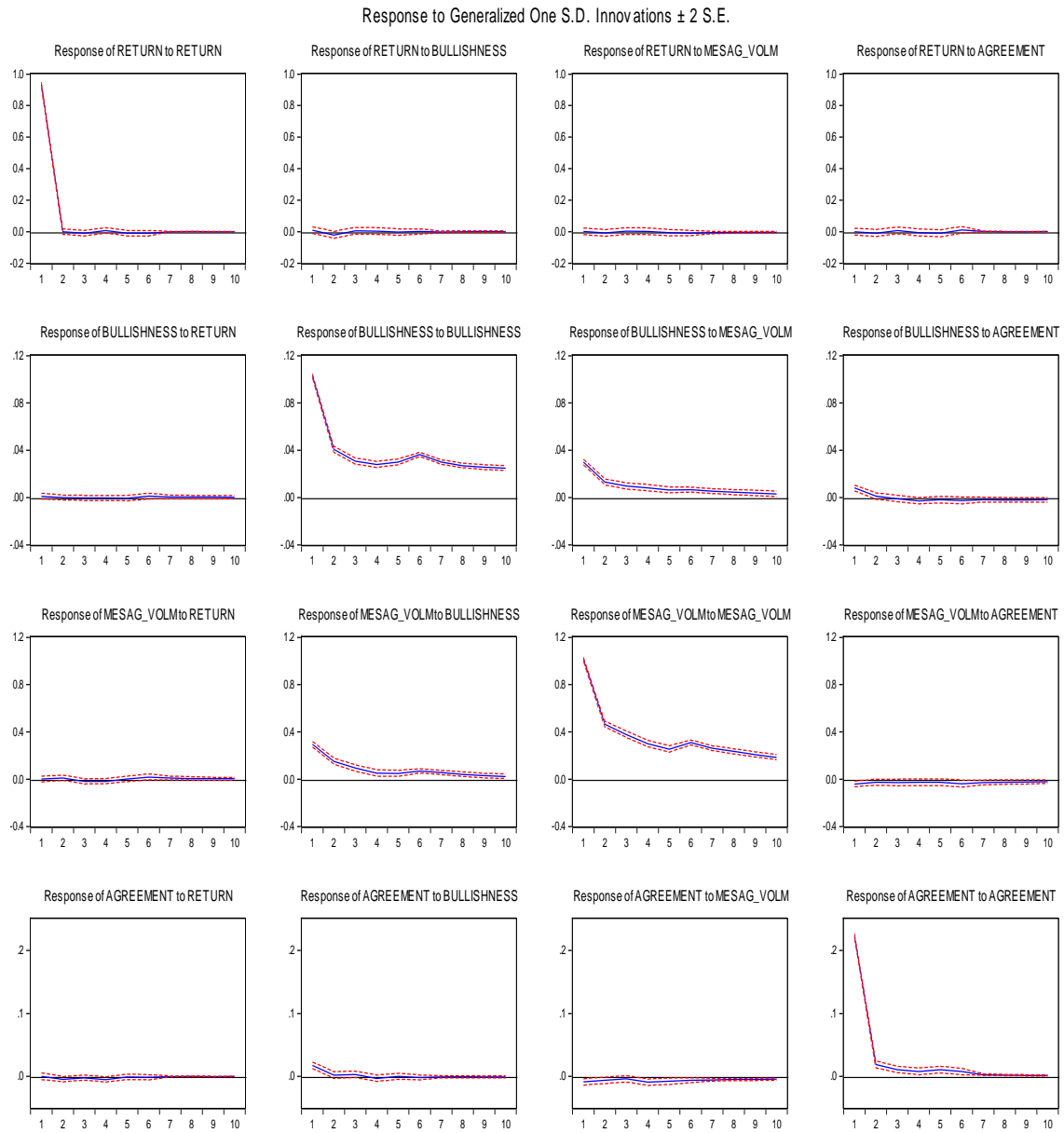


Figure 6.5a: Generalised impulse response functions of short-run Granger causality between stock return and StockTwits variables.

Chapter Six: Empirical Finance Analysis and Discussion

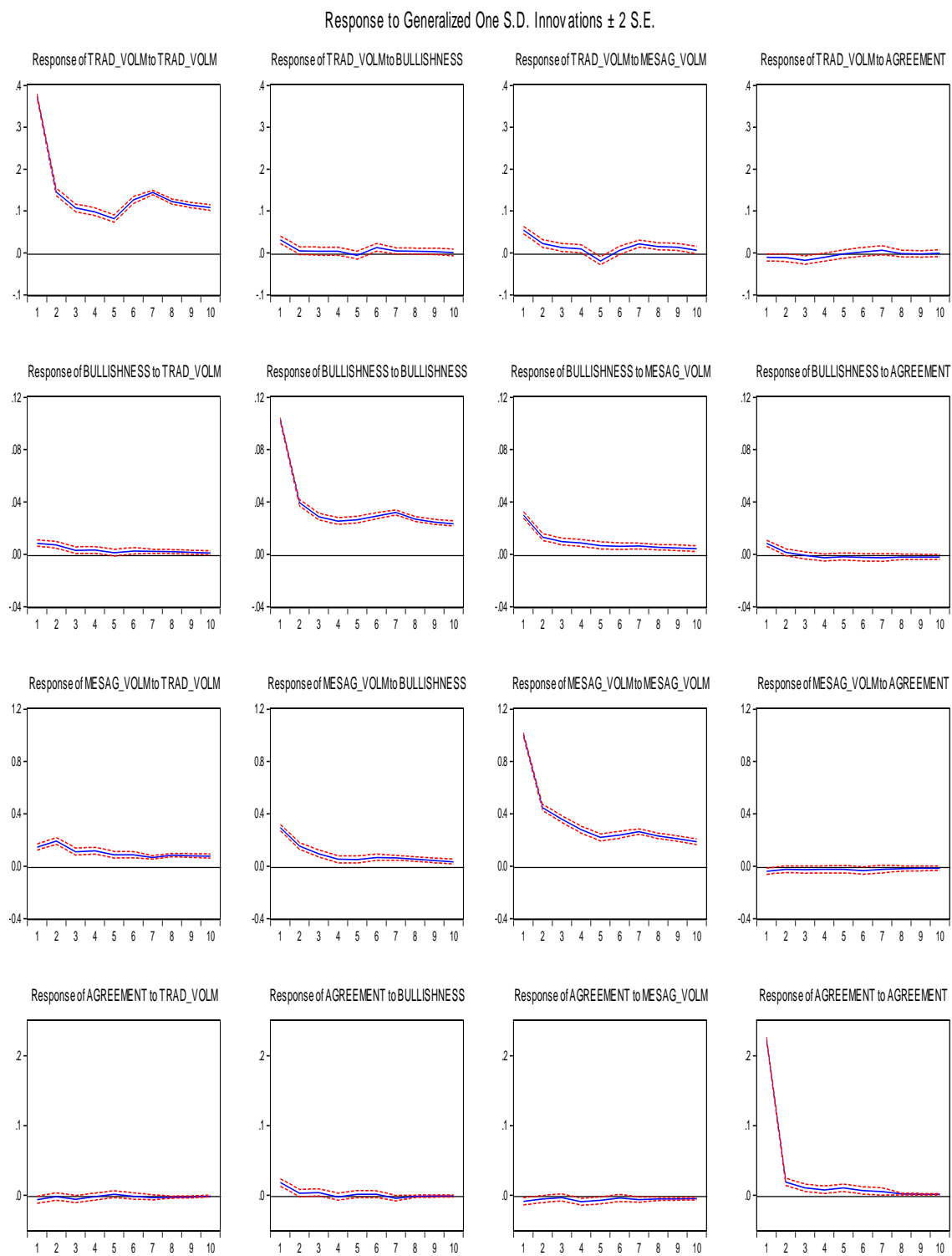


Figure 6.5b: Generalised impulse response functions of short-run Granger causality between trading volume and StockTwits variables.

Chapter Six: Empirical Finance Analysis and Discussion

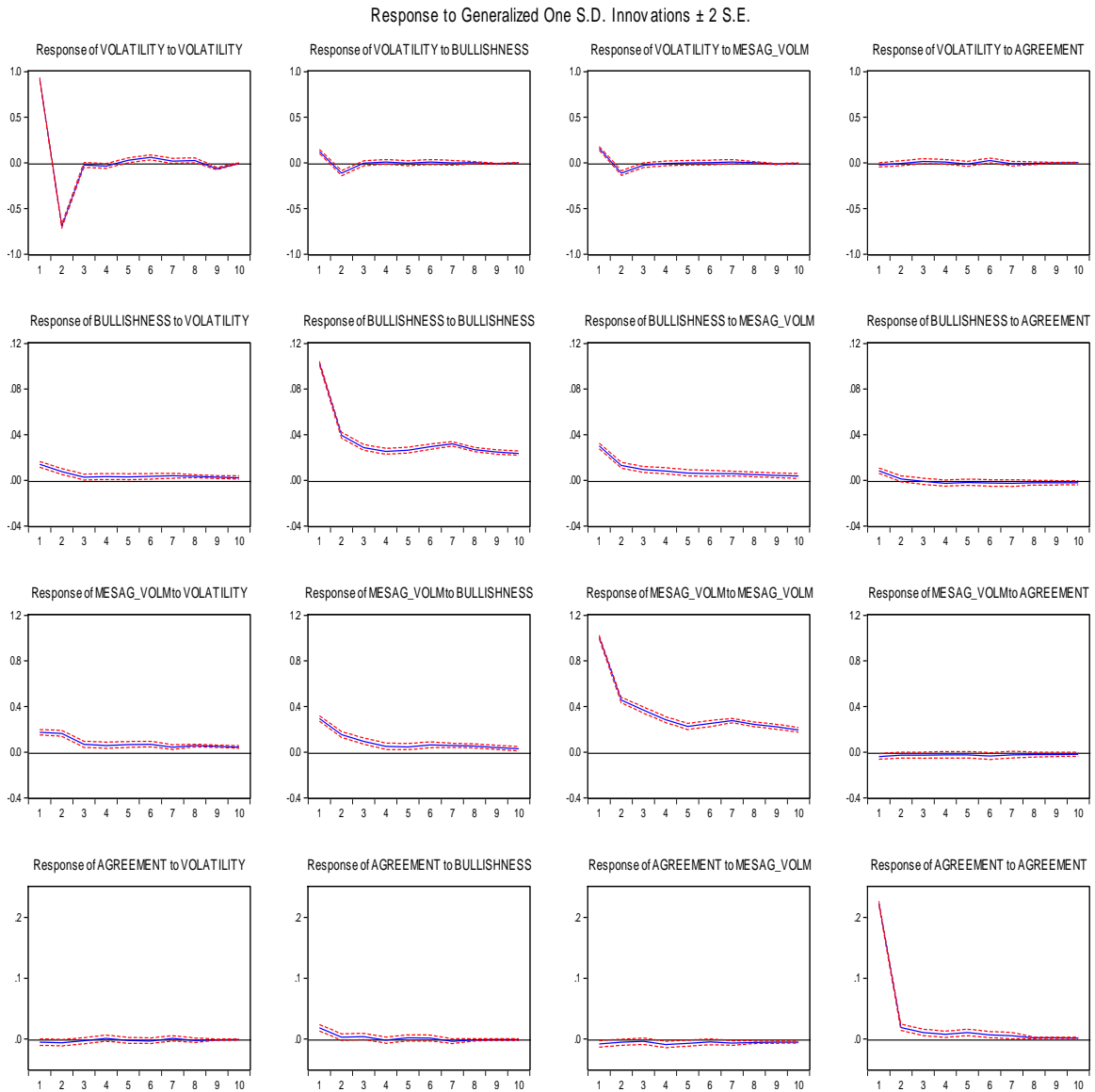


Figure 6.5c: Generalised impulse response functions of short-run Granger causality between trading volume and StockTwits variables

6.7 Chapter Summary

This chapter presented the analysis and findings of the prospective relationships between StockTwits features and financial market indicators. Different correlation analyses, including pairwise correlations, contemporaneous relationships and time-sequencing regressions (VAR models) have been performed to provide a comprehensive and holistic picture of these relationships. The aim was to investigate whether or not StockTwits features contain valuable information that is not yet reflected in stock prices and other financial market indicators. Table 6.10 provides

Chapter Six: Empirical Finance Analysis and Discussion

summary results of the findings of the relationship between stock micro-blogs and financial market activities.

Table 6.10: Summary of the results of the analysis of StockTwits features and stock market

Hypothesis	<i>Contemporaneous Relationship</i>	<i>Lagged Relationship</i>
<i>Message Volume</i>		
H1a Message volume in stock micro-blogging forums has a positive impact on trading volume.	Yes	Yes
H1b Increase in message volume in stock micro-blogging forums is associated with higher stock returns.	No	No
H1c Message volume in stock micro-blogging forums has a positive impact on stock return volatility.	Yes	Yes*
<i>Bullishness (proxy for investor sentiment)</i>		
H2a Investor sentiment derived from stock micro-blogs results in an increase in trading volume.	Yes	Yes
H2b Investor sentiment derived from stock micro-blogs results in an increase in stock market return	No	No
H2c Investor sentiment derived from stock micro-blogs results in an increase in stock return volatility	Yes	Yes
<i>Agreement</i>		
H3a Disagreement among investors in stock micro-blogging forums has a positive impact on trading volume	Yes	Yes
H3b Disagreement among investors in stock micro-blogging forums has a negative impact on stock returns.	No	No
H3c Disagreement among investors in stock micro-blogging forums has a positive impact on stock market volatility	No	Yes

* Note that the relationship between the message volume and volatility tends to negative as indicated in the VAR-volatility model.

To sum up, this research concludes that, for the informants in this study, some StockTwits features appear to contain predictive information with respect to market features (especially bullishness and agreement for trading volume, and message volume and bullishness for volatility) and that StockTwits features have the ability to predict various market indicators. The next chapter elaborates in more detail the relationships between StockTwits features and financial market indicators by considering both the linear and non-linear effects of these relations by empirically investigating whether or not the asymmetric effect of investors' sentiments on the stock market exists.

CHAPTER SEVEN: AN EMPIRICAL INVESTIGATION OF THE ROLE OF INVESTOR SENTIMENT IN THE STOCK MARKET

7.1 Introduction

In developing a deeper understanding of the role played by investor sentiment in predicting stock market behaviour, this chapter investigates the non-linear relationships among the asymmetrical behaviour of investor sentiment in stock market variables by distinguishing between bullish and bearish sentiments. This investigation goes a step further by exploring the asymmetrical responses of investor sentiment to the changes in stock return in different states of the economy (e.g. bull/bear markets). Behavioural finance literature has widely debated whether investors behave differently in different states of the economy. For example, De Bondt (1993) argues that investors' sentiments show extrapolation bias where bullish sentiments are more likely after a period of growth while bearishness is more likely in a period of decline. Verma and Verma (2007) show that innovations in stock markets have a stronger effect on bullish sentiments during a period of positive returns (market growth) than a similar effect on bearish sentiments during a period of negative returns (market recession). The following subsections discuss the issue of the asymmetrical behaviour of investor sentiments in the stock market while showing how these sentiments respond to changes in stock returns in different regimes of the market.

This chapter is divided into six sections including this introduction. Section 7.2 investigates the impact of investor sentiments on stock return, volatility and trading volumes while highlighting the role of noise traders' risk motivated by the DSSW model. This chapter then proceeds to examine the responses of investor sentiments to the change in returns over two stated regimes of the market (bull and bear markets) in section 7.3. Sections 7.4 and 7.5 offer an in-depth analysis of some tweet-market relationships that have created a great puzzle in the empirical literature (e.g. sentiment-return relations and volume-disagreement relations in Sections 7.4 and 7.5 respectively) where more rigorous econometric modelling such as the quantile regression (QR) approach and non-linear modelling are employed. Section 7.6 summarises this chapter.

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

7.2 The Impact of Investor Sentiment on the Stock Market

This section provides empirical investigations of the relationship between investor sentiments, return, volatility and trading volumes. It highlights the critical role noise traders might play in influencing assets prices in capital markets. Additionally, this section investigates asymmetrical behavior of investor sentiment on volatility and trading volume by differentiating between bullish and bearish sentiments.

7.2.1 The Effect of the Change in Investor Sentiment on Stock Return and Volatility (The DSSW (1990) Model)

This thesis investigates the relative influence of rational and noise trading on the stock return and on the formation of return volatility as suggested by DSSW (1990). Since the DSSW (1990) model has long been proven significant in providing substantial evidence of the impacted relations of the change in noise trader sentiment in predicting assets pricing, any empirical test that focuses alone on the impact of sentiment on either the mean return or volatility of the return might provide an incomplete story of these prospective relationships. To tackle this issue, this study follows Lee et al. (2002) by proposing a simple return model that incorporates the effect of noise traders on the formation of volatility and the mean return as suggested in DSSW (1990). More specifically, this research empirically tests the impact of the four effects of the DSSW model namely; the “price pressure” effect, the “hold more” effect, the “Friedman” effect and the “create space” effect as denoted in Figure 7.1. The first two effects have been found to capture the short-term noise traders effects on returns through the inclusions of the contemporaneous shifts in investor sentiments in the return equation. On the other hand, the “Friedman” and the “create space” effects were responsible for the long run associations between noise trader and assets return through the impact of sentiment change on the future volatility.

The noise trader model developed by DSSW (1990) shows that the impact of noise trading on the price of risky assets takes place through the interaction of four effects. The “price pressure” effect states that as noise traders become bullish (bearish), their demand for risky assets creates price pressure that results in purchases (sales) of those assets at prices above (below) their fundamental value and therefore lowers expected returns. On the other hand, the “hold more” effect implies that when

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

the demand by noise traders increases (decreases) relative to their average change in sentiment to become more bullish (bearish), they will expect a higher (lower) return relative to the market risk bearing. These two effects (the price pressure effect and the ‘hold more’ effect) influence stock returns directly, and both account for the direction of shifts in noise trader sentiment indicated by the direct arrows from both effects to the return box in Figure 7.1.

The “Friedman” and “create space” effects are a result of the change in the noise trader’s misperceptions about the asset’s risk. The “Friedman” effect, sometimes called the “buy high-sell low” effect, states that the noise traders tend to buy and sell most stocks simply because other noise traders are buying and selling. As they follow other noise traders by buying and selling when others do so, they are more likely to suffer a capital loss. Moreover, the more variables the noise traders believe in, the more damage their poor market timing does to their returns.

As for the “create space” effect, when the noise traders’ misperceptions about risky assets increase, the price uncertainty of holding those risky assets will also increase, thus reducing the desire of risk-averse arbitrageurs to hold those risky assets. Noise traders are more likely to enjoy higher expected returns by limiting arbitrageurs’ trading activity and deterring them from trading against them. The “Friedman” and “create space” effects are responsible for the magnitude of the shifts in noise traders’ sentiments and indirectly influence stock returns through changes in noise traders’ beliefs/misperceptions of asset risk.

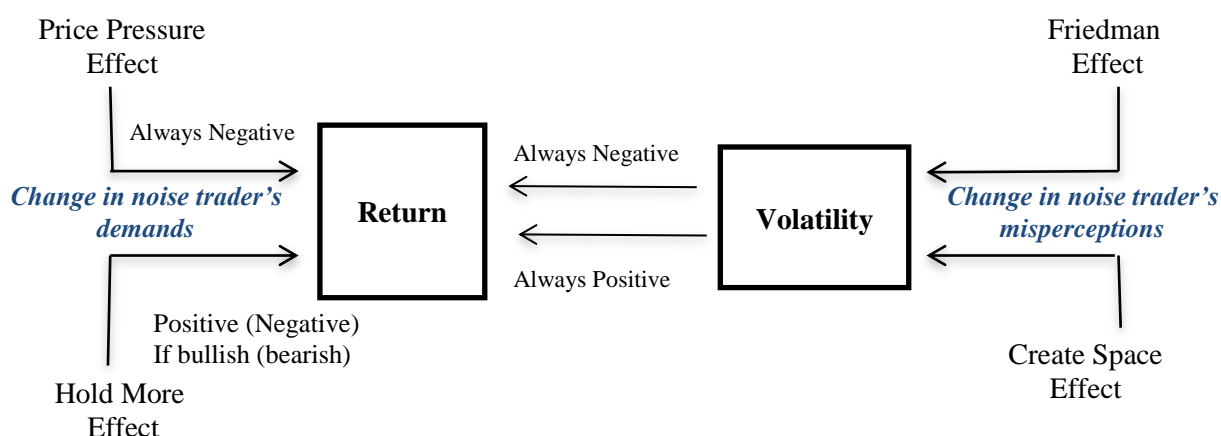


Figure 7.1: The impact of noise trader sentiment on stock returns and volatility
Source: Adopted with modification from Lee et al., (2002).

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

As shown in Figure 7.1, the “Friedman” and the “create space” effects have indirect negative (positive) impact on return through volatility. In general, the “hold more” and “create space” effects tend to increase the expected returns of noise traders while the “price pressure” and “Friedman” effects tend to lower noise traders’ expected returns.

Motivated by the empirical work of Lee et al. (2002), this research examines the effect of investor sentiment on stock returns and volatility by adding the sentiments to mean and variance equations. This section presents the empirical evidence on the relation between sentiment, stock return and volatility by modelling the four effects of noise traders in the mean and volatility equations respectively. To examine the direct impact of noise traders on returns, the contemporaneous shifts in investor sentiment in the return equation are estimated as follows:

$$R_{it} = \alpha_1 + \alpha_2 \Delta v_{it} + \lambda_1 \overline{\Delta B}_{it}^{**} + \beta_1 NWK_t + \beta_2 MKT_t + \varepsilon_{it} \quad (7.1)$$

where R_{it} is the daily return on the i^{th} stocks of the DJIA index, Δv_{it} ³¹ is the daily change in volatility on the i^{th} stocks of DJIA and $\overline{\Delta B}_{it}^{**}$ is a measure of noise traders’ risk associated with the daily shifts in sentiment using the bullishness index extracted from the related stock micro-blogging messages, the so-called “StockTwits”, as a proxy of investor sentiment^{32 33}, MKT is the market return of INDU index (DJIA index return) added to the regressions as a control variable to control for the overall market-wide effects, and NWK is a dummy variable for the first day of the new trading week to control for potential return anomaly effect. In line with Lee et al. (2002), the mean equation is estimated using two alternative measures of noise traders’ risk indicated by model (1) and (2) in Table 7.1. The first measure is computed as the change in the bullishness index, $\overline{\Delta B}_{it}^{**} = \overline{B}_{it}^{**} - \overline{B}_{it-1}^{**}$; while the second measure is computed as the percentage change in the bullishness index of investor, $\% \overline{\Delta B}_{it}^{**} = \frac{\overline{\Delta B}_{it}^{**}}{\overline{B}_{it-1}^{**}}$.

³¹ The Δv_{it} is added to the return regression to reflect the net impact of the DSSW’s “Friedman” effect and “create space” effects in the return. As suggested in the DSSW model that “Friedman” effect and the “create space” effect indirectly influence stock returns through changes in noise traders’ beliefs/misperceptions of asset risk.

³² The bullishness measure serves as a proxy for investor sentiments; therefore, in this research thesis one may use them interchangeably throughout the whole thesis.

³³ In this model adding Δv_{it} in the return equation makes the effects of the change in bullishness $\overline{\Delta B}_{it}^{**}$ in return significant compared to the results founds in quantile regressions as will be seen later in Section 7.4

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

Since the “hold more” and “price pressure” effects both impact the stock return directly, their net effect is reflected in the return equation through the sign and significance of the coefficient λ_1 . If the sign of the coefficient λ_1 is found to be positive as the return tends to be higher when the noise traders become more bullish, this implies that the “hold more” effect dominates. On the other hand, when the net effect is found to be negative, indicated by a negative coefficient of λ_1 , the returns tend to be lower when noise traders become more bearish where both the “price pressure” and “hold more” effects are reinforcing. To clearly demonstrate the lead-lag relationship between investor sentiment and stock return, one period lag of change in sentiment is added to the mean Equation (7.1)³⁴.

The DSSW (1990) theorises that noise traders can affect the volatility of stock returns through the “Friedman” and “create space” effects, hence affecting the stock return indirectly. Therefore, to examine such effects, the volatility equation, which takes the following form, is estimated as follows:

$$\Delta v_{it} = \alpha_1 + \alpha_2 \Delta v_{it-1} + \gamma_1 (\Delta \bar{B}_{it-1}^{**}) D_{it-1} + \gamma_2 (\Delta \bar{B}_{it-1}^{**}) (1 - D_{it-1}) + \beta_1 NWK_t + \beta_2 MKT_t + \varepsilon_{it} \quad (7.2)$$

where D_{it-1} and $(1 - D_{it-1})$ are dummy variables, used to capture the positive (bullish) and negative (bearish) shifts in sentiment for company i (i.e., $i=1, \dots, 30$) at time $t-1$, calculated as follows:

$$D_{it-1} = \begin{cases} 1 & \text{if } \Delta \bar{B}_{it-1}^{**} > 0, \\ 0 & \text{otherwise} \end{cases}, \quad (7.3)$$

$$1 - D_{it-1} = \begin{cases} 1 & \text{if } \Delta \bar{B}_{it-1}^{**} \leq 0. \\ 0 & \text{otherwise} \end{cases}. \quad (7.4)$$

It is expected that the magnitude as well as the direction of shifts in investor sentiment will have an asymmetric impact on stock return volatility. This is because investors perceive the stocks to be more (less) risky and therefore revise their expectation of conditional volatility upwards (downwards). Depending on the types and nature of news, Nelson (1991) finds an asymmetric effect of information arriving

³⁴ Although the results are not reported herein, the lag relationship of change in investor sentiment exert insignificantly effect on stock returns as will be shown later in reported results of Table 7.1.

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

in the market on volatility. Later, Glosten et al. (1993) support Nelson's findings and confirm the asymmetric effect of news on volatility, showing that the magnitude of the effect of bad news on market volatility is greater than the effect of good news. The coefficients (γ_1, γ_2) in the volatility equation (7.2) capture the magnitude effects of shifts in sentiment on volatility. The net effect of the "Friedman" and "create space" effects is captured by and reflected in the estimated coefficient α_2 on the return equation.

From the return regression reported in Table 7.1, it is found that a shift in sentiment has a significantly positive impact on an asset's return only when a change in sentiment $\Delta \bar{B}_{it}^{**}$ is used as a measure of noise trader risk. This means that when the change in investor sentiment $\Delta \bar{B}_{it}^{**}$ increased by 1%, that would lead to a 0.1709 % increase on the daily stock returns of DJIA companies. On the other hand, an insignificant impact of shift in sentiment on return is found when the percentage change in sentiment $\% \Delta \bar{B}_{it}^{**}$ is used. Since the coefficient λ_1 is found to be positive and statistically significant in model (1) when the change in sentiment $\Delta \bar{B}_{it}^{**}$ is used as a measure of noise trader risk, which indicates that an asset's return increases as investors become more bullish. Similarly, Brown and Cliff (2005) find that stock market valuation errors are positively correlated with investor sentiment. This result implies that the "hold more" effect tends to dominate the "price pressure" effect. This means that when noise traders become more bullish about a particular security, their optimism induces traders to hold more of the risky assets than the fundamentals suggest, thereby increasing their expected returns relative to the market risk bearing. However, investors should bear in mind that the higher expected return from holding risky assets may be partially or sometimes fully offset by the unfavorable price caused by the "price pressure" effect through the increased demand to hold more of those risky assets. Similarly, as noise traders become more bearish they tend to be pessimistic and tend to hold fewer risky assets, resulting in them lowering their expected returns and thereby selling off securities.

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

Table 7.1: The relationship between changes in investor sentiment with stock return and volatility

This Table reports the results of the return and volatility models of panel data of DJIA index over the period April 3, 2012 to April 5, 2013. Using the bullishness Index extracted from StockTwits postings as a proxy of investor sentiment, Model 1 and Model 2 incorporate the effect of changes in investor sentiment measured by $\Delta \bar{B}_{it}^{**}$ and $(\% \Delta \bar{B}_{it}^{**})$, respectively. Dummy variables D_{it-1} and $(1 - D_{it-1})$ are used to capture the direction of changes (positive/negative) towards more bullish and bearish sentiments, respectively. Dummy variables are added to capture the first day of the trading week effect to control for Monday return (and or first day after holiday) anomalies. Market returns of DJIA index are added as a control variable to control for market-wide effect. The regressions are estimated based on OLS estimate with company fixed effect.

	<u>Model 1</u>		<u>Model 2</u>	
	Change in Investor Sentiment ($\Delta \bar{B}_{it}^{**}$)		Percentage Change in Investor Sentiment ($\% \Delta \bar{B}_{it}^{**}$)	
<u>Return Regression</u>				
α_1	0.0024	(0.0124)	0.0043	(0.0125)
α_2	-0.0203***	(0.0094)	-0.0176***	(0.0093)
λ_1	0.1709**	(0.0885)	-0.0070	(0.0146)
β_1	0.0123	(0.0250)	0.0082	(0.0250)
β_2	0.9968***	(0.0144)	0.9965***	(0.0144)
R_2	0.3929		0.3926	
N observation	7530		7530	
Durbin Watson stat	1.9915		1.9923	
<u>Volatility Regression</u>				
α_1	0.1121***	(0.0161)	0.0780***	(0.0159)
α_2	-0.4637***	(0.0103)	-0.4646***	(0.0103)
γ_1	-1.0835***	(0.1305)	-0.0923***	(0.0182)
γ_2	1.0524***	(0.1578)	0.3209***	(0.0716)
β_1	-0.1083***	(0.0274)	-0.1100***	(0.0275)
β_2	-0.0795***	(0.0158)	-0.0779	(0.0159)
R^2	0.2256		0.2200	
N observation	7530		7530	
Durbin Watson stat	2.2598		2.2590	

Note (*), (**), and (***) denote significance levels at 10%, 5%, and 1%, respectively. Standard errors are shown in parentheses.

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

It is possible, therefore, to look at the “hold more” effect in terms of the risk return/trade-off concept. In the capital market, the more risk that noise traders can afford to take on by holding more risky assets as their sentiments become more bullish, the higher their expected returns. In contrast, the fewer risky assets that noise traders choose to hold as a result of their bearish sentiments, the lower the returns they may expect from the sale of securities. The finding of a positive correlation between return and shift in sentiments is consistent with Lee et al. (2002) and Verma and Verma (2007), who found that a shift in investor sentiment plays a significant role in determining stock prices in the U.S. market. The results, however, contradict the findings of previous studies that noise trader risk will have an impact only on small traded securities (Lee et al., 1991; Neal and Wheatley, 1998; Wang, 2001; Simon and Wiggins, 2001). A negative relationship between asset returns and stock volatility has been documented in previous studies; i.e. higher (lower) returns are associated with decreases (increases) in stock return volatility.

The lag relationship between shift in investor sentiment and stock return was also tested, although the results is not reported herein, but the results show that the lagged shifts in sentiment in the mean equation do not tend to be statistically significant and their importance is completely captured in the market’s formation of risk in the volatility equation. This is consistent with the findings of Brown (1999) and Lee et al. (2002) who noted that the long-term relationship between sentiment and return is fully reflected in the effect of sentiment in the market formation of risk.

In the volatility Equation (7.2), the findings presented in Table 7.1 reveal that the magnitude of the change and percentage change in sentiment in models (1) and (2) both have a symmetric impact on the formation of assets’ risk. The magnitude of the change and percentage change in sentiment $\Delta \bar{B}_{it-1}^{**}$ and $\% \Delta \bar{B}_{it-1}^{**}$ in models (1) and (2) shows that both bullish and bearish shifts in sentiment are significant in revising the stock volatility. The significant coefficients of γ_1 and γ_2 in models (1) and (2) respectively suggest a volatility spillover from investor sentiments into stock return volatility in the capital market. This study also found an inverse relationship between investor sentiment and stock return volatility. This negative relationship is consistent with earlier findings in previous studies on the negative price of time-varying risk (Glosten et al., 1993; De Santis and Gerard, 1997). More specifically, the study shows that a bullish shift in sentiment, indicated by negative coefficients of γ_1 (e.g., -

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

1.0835 and -0.0923 in model (1) and (2) respectively), resulted in a statistically significant downward revision in volatility. This means that when the bullish shifts in investor sentiment changed on an average of 1%, the daily change in volatility decreased by 1.0835% and 0.0923% as indicated in model (1) and (2) respectively. Conversely, a bearish shift in sentiment may cause an upwards revision in volatility measured by positive coefficients of γ_2 (e.g., +1.0524 and +0.3209) in model (1) and (2) respectively). This means that for a 1% change in bearish investor sentiment, the daily change in volatility is increased by approximately 1.052 % and 0.3209 % as shown in model (1) and (2) respectively. These findings are in line with similar studies by Lee et al. (2002) and Verma and Verma (2007) who found that bullish (bearish) shifts in sentiment may cause a significant downward (upward) revision in conditional volatility.

The positive (negative) effect of bearish (bullish) shifts in sentiment on stock return volatility may explain the interaction of DSSW's "Friedman" and "create space" effects. As stated earlier, the sign and significance of the parameter α_2 in the return regression reflects the net impact of the interactions of DSSW's "Friedman" and "create space" effects on returns. The "Friedman" effect is sometimes called the "buy high-sell low" effect, suggesting that noise traders usually have poor market timing. This implies that increases in noise traders' misperceptions will negatively affect stock prices, thereby lowering noise traders' expected returns. Since both the "Friedman" and "create space" effects reflect the noise traders' impact on the market's formation of risk, their inclusions are more likely to transact together. The extant negative price effect caused by the "Friedman" effect and triggered by the bearish shift in sentiment might be balanced by the space created by noise traders triggered by the bullish sentiment shift. The interactions of these two effects (Friedman and create space) are demonstrated very clearly by the magnitude effect of the coefficients (γ_1, γ_2). Since there is only a slightly greater effect of bullish sentiment shifts, indicated by ($\gamma_1 = -1.0835$) than bearish shifts in sentiment on volatility, indicated by ($\gamma_2 = 1.0524$)³⁵, the higher return associated with the downward revision in volatility caused by bullish sentiment shifts might be enough to offset the lower return associated with poor market timing triggered by bearish shifts

³⁵ To test the magnitude effect of sentiment coefficients (γ_1, γ_2) on return, model (1) that utilised the change in bullishness as a measure of investor sentiment, is used since it shows a significant effect of sentiment on return equation in contrast to model (2) that fail to show any significant impact on return.

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

in sentiment. These results suggest that the “create space” effect may dominate the “Friedman” effect, although the net effect is tiny.

In general, the findings of this research study reveal that bullish (bearish) shifts in investor sentiment cause significant downward (upward) revisions in volatility of returns and are associated with higher (lower) stock returns. These results are consistent with model interactions of noise traders by DSSW (1990) where the permanent effect of noise trading on expected return is captured through its impact on the market formation of risk through the interactions of “Friedman” and “create space” effects. While the direct effect of noise traders on return is captured through the impact of the “hold more” and “price pressure” effects, a positive correlation has been found between return and shift in sentiments, indicating that the higher return associated with the “hold more” effect through the increase in risk premium is relatively greater than the lower return associated with the “price pressure” effect on noise traders. Therefore, it could be concluded that investor sentiment appears to have a predictive power and ability to contribute to the explanation of the assets’ returns while it also showed to be a systematic risk factor that could be measured and priced, which is in line with the noise trader theory of De Long et al. (1990).

7.2.2 The Effect of Investor Sentiment on Trading Volume

In investigating the four effects of the DSSW model of noise traders in the previous section, the results showed that the shift in investor sentiment has a significant role in explaining assets’ returns and plays a tremendous role in formations of assets’ risks in the capital market. Therefore, one might investigate whether or not sentiments also have an impact on volume of trade in the capital market. The assertion of this respective relationship stems from previous empirical studies that find strong evidence that sentiments affect investors’ desire to trade in the stock market. The theories of Black (1986) and Trueman (1988) suggest the persistence of irrational noise traders in the financial market and confirm that noise traders play an important role in providing liquidity, particularly in risky assets (i.e. stocks). Furthermore, Brown (1999) shows that the trading activities of small investors are more pronounced when sentiments are at the extreme level (extremely high/low). More

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

recently, Yuan (2008) and Karlsson et al. (2009) show that sentiment-driven investors participate and trade more aggressively when sentiments are high.

This section aims to provide a better understanding of the relationship between sentiment and trading volumes. Understanding such a relationship may help us to assess the impact of noise traders. Therefore, to investigate the effect of sentiments on trading volume, the volume regression is estimated on the two sentiment measures, as previously explained in the mean equation (Eq.7.1) in Section 7.2.1. These two sentiment measures are the change in bullishness and the percentage change in bullishness. The study then further investigates the existence of an asymmetric effect of shift in sentiment on trading volume by differentiating between bullish and bearish sentiments. In examining the asymmetric impact of bullish and bearish shifts in sentiments, the empirical investigations in this thesis will also consider both the change and the percentage change in sentiments.

$$TV_{it} = \alpha_1 + \phi TV_{it-1} + \lambda_1 \Delta \bar{B}_{it}^{**} + \beta_1 NWK_t + \beta_2 MKT_t + \varepsilon_{it} \quad (7.5)$$

where, TV_{it} is the daily trading volume on the i^{th} stocks of the DJIA index, $\Delta \bar{B}_{it}^{**}$ is the daily shifts in sentiment using the bullishness measure, TV_{it-1} is the lagged trading volume at time $t-1$ ³⁶³⁷, NWK and MKT is the new day of the week dummy and market index are added, respectively, to the volume regression to control for the new day of the trading week or first day after holiday dummy and for market wide effects. A panel regression with company fixed effect is used to estimate the volume regression in Eq. (7.5). The volume regression is estimated twice in two models (model (1) and model (2)) each with different measures of sentiment. The change in bullishness is utilised in model 1 and the percentage change in bullishness is utilised in model 2.

It is apparent from Table 7.2 that investor sentiment contains new information that is not yet reflected in volume of trade. Both models resulted in a very high R^2 of 0.85 and 0.84 in model (1) and (2), respectively, indicating the significant

³⁶One period lag of trading volume is added to the volume regression equation (7.5) to overcome with the problem of serial correlation in the model equations.

³⁷In all of our empirical results in this chapter, the residuals are checked to confirm that it is free of serial correlation on the bases of Ljung-Box test.

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

improvement in the goodness of fit when either the change in bullishness or the percentage change in bullishness is included as an explanatory variable in OLS trading volume models (1) and (2).

Table 7.2: The relationship between changes in investor sentiment and trading volume

This Table reports the results of the trading volume model of panel data of DJIA index over the period April 3, 2012 to April 5, 2013. Using the bullishness Index extracted from StockTwits postings as a proxy of investor sentiment, Model 1 and Model 2 incorporate the effect of changes in investor sentiment measured by $\Delta\bar{B}_{it}^{**}$ and $(\%\Delta\bar{B}_{it}^{**})$, respectively. Dummy variables are added to capture first day of the trading week effect to control for Monday return (or first day after holiday) anomaly. Market returns of DJIA index are added as a control variable to control for market-wide effect. The regressions are estimated based on OLS estimate with company fixed effect.

	<u>Model 1</u>	<u>Model 2</u>
	Change in Investor Sentiment $(\Delta\bar{B}_{it}^{**})$	Percentage Change in Investor Sentiment $(\%\Delta\bar{B}_{it}^{**})$
<u>Trading Volume Regression</u>		
α_1	9.7051*** (0.1599)	9.7552*** (0.1603)
ϕ	0.3377*** (0.0110)	0.3342*** (0.0110)
λ_1	0.1168*** (0.0393)	0.0248*** (0.0057)
β_1	-0.1497*** (0.0099)	-0.1526*** (0.0099)
β_2	-0.0221*** (0.0056)	-0.0221*** (0.0056)
R^2	0.8451	0.8446
N observation	7530	7530
Durbin Watson	2.0889	2.0841

Note (*), (**), and (***) denote significance levels at 10%, 5%, and 1%, respectively. Standard errors are shown in parentheses.

The shift in investor sentiment shows a significantly positive and an economically important impact on trading volume regardless of the measure of sentiment used, namely the change in bullishness ($\Delta\bar{B}_{it}^{**}$) or the percentage change in bullishness ($\%\Delta\bar{B}_{it}^{**}$). In model 1, for a 1% change in investor sentiment, the daily trading volume is increased by approximately 0.1168% (indicated by $\lambda_1 = +0.1168$). A possible explanation for these results is that when bullishness of investors (measured by the $\Delta\bar{B}_{it}^{**}$) increased on average by one bullish investor, the current trading

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

volumes would increase by 0.1168% of the normal traded shares on that given day. Similarly, in model 2, for a 1% change in sentiment, there is approximately a 0.0248% increase in the current daily trading volumes (indicated by $\lambda_1 = +0.0248$). This means that when bullishness of investors (measured by the $\% \Delta \bar{B}_{it}^{**}$) increased on average by one bullish investor, the current trading volumes would increase by 0.0248% of the normal traded shares on that given day. Although the change in investor sentiment shows a stronger effect on trading volume than a percentage change in sentiment, both show a very significant and positive effect on trading volume. This result implies that when noise traders become more bullish they are more likely to translate their optimism through the buying of more shares, thus facilitating liquidity in the market. This provides support for the “trend chasing” behaviour demonstrated by Kurov (2008), where noise traders tend to trade more actively during bullish/high sentiment periods, thereby increasing market liquidity. Benos (1998) and Odean (1998b), in testing the over-confidence theory of excessive trading, have demonstrated that, due to over-confidence, investors tend to trade too much.

After demonstrating the significant effect of the shift in investor sentiments on trading volume, this research then examines the existence of the asymmetrical behaviour effect of investor sentiments by considering the effect of change in bullish and bearish shifts in sentiments on trading volume. The following model incorporates the asymmetric effect of bullish and bearish shifts in sentiments on trading volume as follows:

$$TV_{it} = \alpha_1 + \phi TV_{it-1} + \gamma_1 (\Delta \bar{B}_{it}^{**}) D_{it} + \gamma_2 (\Delta \bar{B}_{it}^{**}) (1 - D_{it}) + \beta_1 NWK_t + \beta_2 MKT_t + \varepsilon_{it} \quad (7.6)$$

where, TV_{it} is the trading volume of stock i on day t , $\Delta \bar{B}_{it}^{**}$ is the change in investor bullishness (serving as a proxy for investor sentiment) at time t , D_{it} is the dummy variable that captures the positive change in investor sentiment, and $(1 - D_{it})$ is the dummy variable that represents the negative change. $(\Delta \bar{B}_{it}^{**}) D_{it}$ and $(\Delta \bar{B}_{it}^{**}) (1 - D_{it})$ are both interaction terms representing the bullish and bearish shift in sentiment, respectively. The model in Eq. (7.6) suffers from serial correlation therefore one

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

period lag of trading volume is added to remove the serial correlation problem. This study estimates the trading volume equations in (7.6) using contemporaneous panel regression with company fixed effect employing both sentiment measures (the change and the percentage change in sentiments) separately in models (1) and (2) at time t .³⁸

Table 7.3: The asymmetric impact of bullish and bearish shift in sentiment on trading volume

This table reports the results of the asymmetric effect of the bullish and bearish shifts in sentiment on trading volume by using change in investor sentiment $\Delta \bar{B}_{it}^{**}$ as a measure of noise trader sentiment. Model 1 and Model 2 incorporate the effect of changes in investor sentiment measured by $\Delta \bar{B}_{it}^{**}$ and $(\% \Delta \bar{B}_{it}^{**})$, Dummy variables D_{it} and $(1 - D_{it})$ are used to capture the direction of changes (positive/negative) towards more bullish and bearish sentiments, respectively. The regressions are estimated based on OLS estimates with company fixed effect.

	<u>Model 1</u>		<u>Model 2</u>	
	Change in Investor Sentiment ($\Delta \bar{B}_{it}^{**}$)		Percentage Change in Investor Sentiment ($\% \Delta \bar{B}_{it}^{**}$)	
<u>Trading Volume Regression</u>				
α_1	9.8555***	(0.1608)	9.7759***	(0.0060)
ϕ	0.3259***	(0.0110)	0.3325***	(0.0111)
γ_1	0.4507***	(0.0470)	0.0257 ***	(0.0064)
γ_2	-0.1256***	(0.0561)	-0.0273	(0.0254)
β_1	-0.1503***	(0.0099)	-0.1534***	(0.0099)
β_2	-0.0227 ***	(0.0056)	-0.0223***	(0.0056)
R^2	0.8461		0.8445	
N observations	7530		7530	
Durbin Watson stat	2.0827		2.0832	

Note (*), (**), and (***) denote significance levels at 10%, 5%, and 1%, respectively. Standard errors are shown in parentheses.

This research investigates the possibility of an asymmetric impact of bullish and bearish shifts in sentiments on trading volume by examining the coefficients γ_1 and γ_2 . As shown in Table 7.3, the resulting output of the estimated volume equation in model (1) reveals that the coefficients γ_1 and γ_2 for the bullish and bearish shift respectively are highly and statistically significant. It is found that the bullish shifts in investor sentiment are strongly positively correlated with the volume of trade

³⁸ Although the result is not reported herein, it is found that the lag shift in bullish and bearish sentiment at time $t-1$ (using the change in investor sentiment as in model 1) also confirms the existence of the asymmetric effect of sentiments on trading volume (whereby bullish (bearish) sentiment triggered increases (decreases) in volume of trade). This study only reports the asymmetric effect with contemporaneous effect.

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

indicated by the significant positive coefficient of ($\gamma_1 = +0.4507$, p-value < 0.01) whereas the bearish shifts is found to be negatively correlated with trading volume measured by the negative coefficient of ($\gamma_2 = -0.1256$, p-value < 0.001). This result suggests a significant positive impact of investor sentiment on stock trading volume; i.e. the volume of trade increases (decreases) when investors become more bullish (bearish). The estimated coefficient of bullish /bearish shift in sentiments in model (1) could be interpreted as follow: for a 1% change in bullish sentiments, there is approximately a 0.4507% increase in the current daily trading volumes, whereas, a 1% change in bearish sentiments would lead to a 0.1256% decrease in the current volumes of traded stocks. Model (2) by contrast shows that only bullish shifts in sentiments in the current period result in statistical significant increase in trading volumes of the same period indicated by the positive coefficient of ($\gamma_1 = +0.0257$, p-value < 0.01) whereas bearish shifts in sentiments shows insignificant results. Since only a bullish shift in sentiment shows a significant result, its estimated coefficient is interpreted as a 1% change in bullish sentiments of investors, generate 0.0257% increase in the current trading volumes of DJIA stocks.

A theoretical model concerning the direct linkage between sentiments and trading volume is quite limited³⁹, while there is a considerable body of empirical literature on investor trading behaviour and market stability (i.e. Lakonishok et al., 1992; Kamara et al., 1992; Wermers, 1999). Although De Long et al. (1990) do not explicitly model trading volume, their model shows how prices diverge significantly from fundamental values, thus causing noise traders and rational investors to speculate and engage in trading and enabling noise traders to earn higher expected returns than arbitrageurs who trade against them. The DSSW model shows that the volume reactions are produced by the underlying relations of the risk-return to investors in the capital market. Later empirical studies show that traders are destabilising with changes in market price stability and that traders are positive feedback traders - they buy past winner stocks and sell past losers (DSSW, 1990; Lakonishok et al., 1992; Kurov, 2008). In other words, positive feedback trading implies that investors trade less if past returns were negative and trade more if past

³⁹ Baker and Stein (2004) used market liquidity as measured by trading volume as an indicator of investor sentiments. They demonstrate that the trading volume contains information on investor sentiment and that an increase in trading volume reflects a rise in investor sentiments. Although their study utilised trading volume as a sentiments indicator, their focus was on how trading volume as a sentiment measure will have a valuation effect on stocks rather than studying the effect of sentiments on volume of trade.

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

returns were positive. These previous studies show that increases in bullishness are triggered by past positive returns in the bull market whereas bearishness is triggered by past negative returns in the bear market. Therefore, if investors are enthusiastic (bullish investors) about the future movement of stock prices they are more likely to trade more actively, but if they hold pessimistic beliefs about the future development of the stock market, they may trade less. This argument is in line with the results shown in Table 7.3 which reveals that bullish shifts in investor sentiments trigger an increase in the volume of traded securities while bearish shifts in sentiments trigger a reduction in the volume of traded securities in the capital market.

In light of the risk- return trade off theory, investors with high levels of optimism tend to hold more risky assets, which will increase the level of market risk and thereby result in a higher expected return; this causes them to buy more of the winning securities, thereby increasing the trading volume of traded securities. On the other hand, the results show a reduction in excess returns as sentiments become more bearish. These findings confirm the noise trader theory (DSSW, 1990) that the likelihood of holding risky assets decreases when noise traders become more pessimistic and that negative price impact caused by sentiment induces sales of securities, causing a decline in trading volume of traded securities. These results concur with Baker and Stein (2004) and Barber and Odean (2008) who demonstrate that, in the presence of short-sell constraints, bullish noise traders tend to buy more shares, thus boosting liquidity. In contrast, if noise traders are bearish about the market, short-sales constraints, which render them reluctant to take a short position, will therefore keep them out of the market.

Specifically, the results reveal that bullish shifts in sentiments have a greater effect on security trading volume than do bearish shifts in sentiments. This result is consistent with the investor sentiment model developed by DSSW (1990) who suggest that the optimism and pessimism of noise traders, with their erroneous stochastic beliefs, have the power to change the prices, causing a transitory divergence between prices and fundamental values and thus inducing trading. This also supports other behavioural explanations (Brown and Cliff, 2005; Gervais and Odean, 2001; Wang, 2001; Hong et al., 2000) that claim that the effect of sentiments on stock prices can be asymmetric.

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

7.3 Investor Sentiment Reactions to Different Regimes of the Market (Bull and Bear Markets)

The stock market is driven by news. It has been argued that the effect of good news on the market differs from that of bad news. This is called the asymmetric effect, in that the good news does not boost the market as much as the bad news dampens it. This brief argument, however, leads us to the most widely discussed question in the literature about whether market participants react more harshly if bad news is disseminated. In the context of this research study, the good news is defined as an upward market movement (market exhibits positive return), which may be called a bull market, whereas the bad news is defined as a downward market movement (market exhibits negative return), which may be called a bear market. It has been widely argued that the existence of noise traders and arbitrageurs and their trading behaviour in the capital market are greatly affected by their sentiments. The changes in investor sentiment as either bullish or bearish are related to the direction of shifts in asset returns as either positive or negative. For example, if noise traders experience a positive increase (decrease) in returns (as good/bad news arrives) of a given stock in the market, they are more likely to become optimistic (pessimistic) toward that particular security, thus supporting the positive feedback trading theory (De Long et al., 1990). Therefore the possibility of an asymmetric impact of return in the bull and bear market on investor bullishness is investigated by testing the hypothesis that negative market returns (bear market) influence investor bullishness more than positive returns (bull market).

The bullishness equation is estimated by including two market regimes (bull and bear market) to show whether the contemporaneous returns help to explain investor sentiment. Therefore, the model employed in this section is used to statistically identify two different regimes/patterns of return: bull market and bear market. The aim is to distinguish the impact of return on investor sentiments in these two regimes individually. To implement this test, two indicator variables are created and labelled I^{bull} and I^{bear} for the bull and bear market effect respectively. To reflect the significance of returns in the bullishness equation, the bull and bear dummies are created and defined as follows:

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

$$I_{it}^{bull} = \begin{cases} 1 & \text{if } r_{it} > 0 \\ 0 & \text{otherwise} \end{cases} \quad (7.7)$$

$$I_{it}^{bear} = \begin{cases} 1 & \text{if } r_{it} \leq 0 \\ 0 & \text{otherwise} \end{cases} \quad (7.8)$$

An interaction term of each indicator variable with contemporaneous market returns $I_{it}^{bull} r_{it}$ and $I_{it}^{bear} r_{it}$ is added in the bullishness Eq. (7.9) for the bull and bear markets respectively⁴⁰. This procedure allows for an asymmetric response of investor bullishness to market returns in the bull and bear markets accordingly as follows:

$$\bar{B}_{it}^{**} = \alpha_1 + \phi \bar{B}_{it-1}^{**} + \gamma_1 I_{it}^{bull} r_{it} + \gamma_2 I_{it}^{bear} r_{it} + \beta_1 NWK_t + \beta_2 MKT_t + \varepsilon_{it}, \quad (7.9)$$

The results presented in Table 7.4 indicate that the model specification in Equation (7.9) suffers from serial autocorrelation in the data series. Therefore, the model is modified to include lagged bullishness to assist in removing the serial correlations (Dickey and Fuller, 1979; Balvers et al., 2000). The panel regression with company fixed effects is used where the market index and first-day-of-the week dummy were added to the regression to control for the market-wide effect and the negative return on the first trading day of the week, respectively.

The results reported in Table 7.4 show that bullishness (as a proxy for investor sentiment) tends to respond to stock returns positively in the bull market and negatively in the bear market. The coefficients γ_1 and γ_2 measure the investor sentiment response to market return news in bull and bear markets, respectively. It is found that both interaction terms' parameters are statistically significant at the 1% level of significance. The significant positive coefficient of $\gamma_1 = +0.0157$ indicates that positive returns trigger an increase in investor bullishness (decrease in bearishness) in the bull market by 1.57 %, while the negative coefficient of $\gamma_2 = -0.0122$ implies therefore that a negative return triggers a reduction in investor bullishness by 1.22% and or (increase bearishness by 1.22%) in the bear market. A possible explanation for

⁴⁰ Although the results are not reported herein, it is found that when both changes in investor sentiment measured by $\Delta \bar{B}_{it}^{**}$ and $(\% \Delta \bar{B}_{it}^{**})$, are used as a dependent variable in Eq. (7.9), similar statistical results are achieved. Both the change $\Delta \bar{B}_{it}^{**}$ and the percentage change $\% \Delta \bar{B}_{it}^{**}$ in sentiment tend to respond to stock return positively (negatively) in the bull (bear) market. In this thesis, bullishness level is used as a dependent variable in Eq. (7.9) because using this model seems to fit the data very well indicated by high R^2 of 62.9% (a measure of goodness of fit) compared to the other two sentiment measures that results in a very tiny percentage of R^2 where the validity of these model will be in question.

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

this is that, when the stock return r_t is positive (negative), investor bullishness exhibits a pronounced increase (decrease), which implies that in the bull market an investor becomes more bullish whereas in the bear market investor is likely become more bearish. These findings support early empirical literature (Odean 1998b, and then Gervais and Odean, 2001) that demonstrates that the overconfidence of noise traders increases as they attribute high return in bull market. These findings further support the subjective evidence: When the market is on a bull run as it was in the late 1990s, investors appear to become more bullish. This finding is consistent with (DeBondt, 1993) who found that increased bullishness could be expected after a market rise and increased bearishness after a market fall. These findings suggest that stock returns and investor sentiment can act as a system and imply the positive role of the stock market in the formation of investor sentiment (Verma and Verma, 2007). This evidence is also in line with the existence of bandwagon effect (Brown and Cliff, 2004), which states that good returns in a given period drive optimism and they found that stock returns predict sentiments.

The strength of the effect of positive return (bull market) and negative return (bear market) to the bullishness can be measured however by the magnitude of the parameters of γ_1 and γ_2 . The magnitude of the impact of returns on bullishness appears to be greater in the case of the positive returns (in the bull market) compared to decreasing stock prices (bear market). This finding is in line with Verma and Verma (2007) who found a stronger effect on bullish sentiments during the period of positive return (growth) than the effects on bearishness during the period of negative return (decline). The results of this study contradict the findings of Kling and Gao (2008) who show that negative returns have a stronger impact on investor sentiments (decrease in bullishness) than positive returns (increase in bullishness). In general, the findings confirm the existence of an asymmetric effect of stock returns on the bullishness of investors since there is a greater positive impact on bullishness during the period of growth in the bull market than the negative impact on bullishness during the period of decline in the bear market.

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

Table 7.4: The asymmetric response of the investor sentiment to the change in stock returns in the bull and bear markets.

This table report the estimated coefficients of the following regression:

$$\bar{B}_{it}^{**} = \alpha_1 + \phi \bar{B}_{it-1}^{**} + \gamma_1 I_{it}^{bull} r_{it} + \gamma_2 I_{it}^{bear} r_{it} + \beta_1 NWK_t + \beta_2 MKT_t + \varepsilon_{it},$$

where, I^{bull} and I^{bear} are two dummy variables that capture the market growth (bull market) and the market recession (bear market), respectively. The regressions are estimated based on OLS panel regression with company fixed effect.

Bullishness Regression	Contemporaneous regression at time (t)
α_1	0.1885*** (0.0038)
ϕ	0.3592*** (0.0107)
γ_1	0.0157*** (0.0019)
γ_2	-0.0122*** (0.0019)
β_1	-0.0164*** (0.0027)
β_2	-0.0037* (0.0020)
R^2	0.6290
N-observations	7,530
Durbin-Watson statistics	2.0845

Note (*), (**), and (***) denote significance levels at 10%, 5%, and 1%, respectively. Standard errors are shown in parentheses.

7.4 Dispersion of Stock Returns and Investor Sentiment: A Quantile Regression Approach

This section examines the impact of investor sentiment on stock return based on quantile regressions. In addition to investigating the effect of sentiment level on stock return (As in Chapter 6), this study also investigates the effect of asymmetrical behaviour of investor sentiment by distinguishing between bullish and bearish sentiments on stock return. The aim is to provide a comprehensive description of the effect of sentiment cross a range of quantiles of the conditional return distribution. This enables to study the behaviour of extreme quantiles associated with large positive and negative returns rather than the central quantile that is equivalent to the conditional mean in the ordinary least-squares regressions.

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

7.4.1 Empirical Methodology and Model Specifications

Departing from previous studies that confine the analysis to the conditional mean of return distribution to examine the influence of investor sentiment on return, this study revisits the relationship between investor sentiment and stock returns using the quantile regression (QR) technique. Moreover, it investigates the effect of asymmetrical behavior of investor sentiments on stock return by differentiating between bullish and bearish sentiment. This research uses the quantile regression framework as introduced by Koenker and Bassett (1978) and examines the influence of the shift in bullishness on all quantiles of the current return. One of the main attractions of employing the quantile regression in this research is that given the stylised fact that the financial returns are not normally distributed; the QR has several advantages, which, in turn, can address some of the potential pitfalls of earlier studies. First, the QR technique provides more robust results of the coefficient estimates compared to those obtained by the OLS, since such a model is unresponsive to the effect of the outliers in the data and also to the non-normal distribution feature of the error terms. Second, as documented in Chevapatrakul (2015) among others, an assumption with regard to the distribution of the error term is not required, given the semi parametric nature of the QR. Third, the QR can also be used to detect the presence of asymmetry of the slope parameter (capturing bullish and bearish sentiment in case of this study), computed at the various quantiles.⁴¹As shown by Feng et al. (2008), Ma and Pohlman (2008), Chuang et al. (2008), Baur et al. (2012), Alagidede and Panagiotidis (2012), and Chevapatrakul (2015) among others, the QR has been particularly appropriate for modelling stock returns. The purpose of the study is to model the quantile of stock return for a given bullishness level based on a linear model as well as considering the asymmetric non linear behaviour of investor sentiment (bullish and bearish sentiment) on stock return. In many cases, quantile regression estimates are quite different from OLS models. These results carry crucial implications for the linkage between investor sentiment and stock markets.

- **The OLS model**

The primary aim of this research is to examine the impact of investor sentiment on stock returns. The benchmark model commonly used in the literature is

⁴¹ For detailed advantages of the QR when modelling stock returns, see also Chevapatrakul (2015).

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

specified as follows:

$$R_{it} = \alpha + \beta \Delta \bar{B}_{it}^{**} + \gamma_1 MKT_t + \gamma_2 NWK_t + \varepsilon_{it}, \quad (7.10)$$

where R_{it} is the daily stock return for company i at time t , $\Delta \bar{B}_{it}^{**}$ indicates the change bullishness measure, which proxies investor sentiment, for company i at time t ,⁴² MKT is the market return of INDU index (DJIA index return) added to the regressions as a control variable to control for the overall market-wide effects, NWK is a dummy variable for the first day of the new trading week to control for potential return anomaly effects and ε_{it} is the error term.

The general argument of the sentiment-return nexus in the classical finance literature is that sentiment plays no role in affecting stock returns, suggesting that stock prices follow a random walk. The behavioural finance theory, in contrast, posits the existence of two heterogeneous agents, namely noise traders and arbitrageurs. Moreover, the effects of their trading behaviour have the power to affect stock prices, thereby suggesting the existence of some degree of predictability (De Long et al., 1990). This implies that irrational sentiments of optimism or pessimism can affect stock prices for a significant period of time.

The recent literature has also gone a step further by investigating the asymmetric effect of bullish and bearish sentiments on stock returns (see, among others, Lee et al., 2002; Verma and Verma, 2007). In order to examine such an effect using StockTwits data, the model is specified as follows:

$$R_{it} = \alpha + \beta_1 \Delta \bar{B}_{it}^{**} D_{it} + \beta_2 \Delta \bar{B}_{it}^{**} (1 - D_{it}) + \gamma_1 MKT_t + \gamma_2 (\mp) NWK_t + \varepsilon_{it} \quad (7.11)$$

where D_{it} and $(1 - D_{it})$ are dummy variables, used to capture the positive (bullish) and negative (bearish) shifts in sentiment for company i (i.e., $i=1, \dots, 30$) at time t . As previously defined and explained in Section 7.2.1 by the Eq. (7.3) and (7.4) for the bullish and bearish sentiments respectively, the dummy variables in Eq. (7.11) however are created at time t rather than $t-1$. That is D_{it} takes value of one if $\Delta \bar{B}_{it}^{**} > 0$

⁴² In line with (Lee et al., 2002) where the change in investor sentiment is computed as, $\Delta S_t = S_t - S_{t-1}$ likewise we compute the change in bullishness index which serves as investor sentiment proxy as follows; $\Delta \bar{B}_{it}^{**} = \bar{B}_{it}^{**} - \bar{B}_{it-1}^{**}$.

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

and zero otherwise, whereas $(1 - D_{it})$ takes value of one if $\Delta\bar{B}_{it}^{**} \leq 0$; and zero otherwise.

- **Quantile Regression**

This study revisited the relationship between investor sentiment and stock returns by using the QR technique, developed by Koenker and Bassett (1978). It considers the following conditional quantile model:

$$R_{it} = \alpha(\tau) + \beta(\tau)\Delta\bar{B}_{it}^{**} + \gamma_1(\tau)MKT_t + \gamma_2(\tau)NWK_t + \varepsilon_{it}(\tau) = x'_{it}\boldsymbol{\theta}(\tau) + \varepsilon_{it}(\tau), \quad (7.12)$$

where τ denotes the τ -th conditional quantile of stock i 's return, and, $\Delta\bar{B}_{it}^{**}$ is the shift in bullishness of the corresponding stocks, MKT and NWK are the market control variable and the first day of the week dummy respectively, as defined earlier. The estimated coefficient of $\beta(\tau)$ is the main concern in this model specification, which can be interpreted as a parameter estimate of a specific τ -th conditional quantile. Moreover, to uncover the asymmetric effect of sentiment on stock returns, Eq. (7.12) is re-specified by including the shift in bullish and bearish sentiments separately in the model as follows:

$$R_{it} = \alpha(\tau) + \beta_1(\tau)\Delta\bar{B}_{it}^{**}D_{it} + \beta_2(\tau)\Delta\bar{B}_{it}^{**}(1 - D_{it}) + \gamma_1(\tau)MKT_t + \gamma_2(\tau)NWK_t + \varepsilon_{it}(\tau) = z'_{it}\boldsymbol{\psi}(\tau) + \varepsilon_{it}(\tau), \quad (7.13)$$

The aim of this model specification is to assess the influence of both bullish and bearish shifts in sentiment on the different conditional quantiles of stock returns measured by $\beta_i(\tau)$, $i = 1$ and 2 .

Moreover, it is worth noting that Eqs. (7.12) and (7.13) are estimated with 9 quantiles (i.e., $\tau = 0.05, 0.1, 0.25 \dots 0.95$). The entire distribution of the regressor is traced conditional on the regress and as τ is increased from 0 to 1. The 9 quantiles are further divided into three different quantile levels: low, medium and high. The rule of thumb followed in this research study is that a quantile level exerts a statistically significant influence if there are at least two adjacent quantiles that are statistically

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

significant in that corresponding quantile's level. The standard errors are obtained using the bootstrap method.

7.4.2 Empirical Results

This Section provides estimates of the linear sentiment effects on stock returns using the OLS and QR method. Then, the asymmetric effects of shifts in investor sentiment, considered by distinguishing between bullish and bearish sentiments, are reported and discussed.

- **Linear model results**

The parameter estimates of $\beta(\tau)$ and their 95% confidence intervals (in the shaded area) against τ along with the OLS estimate (dashed line) and its 95% confidence interval (dotted line) are plotted in Figure 7.2. The corresponding numerical results of the OLS allowing for company fixed effects and those of the quantile regressions are reported in Table 7.5.

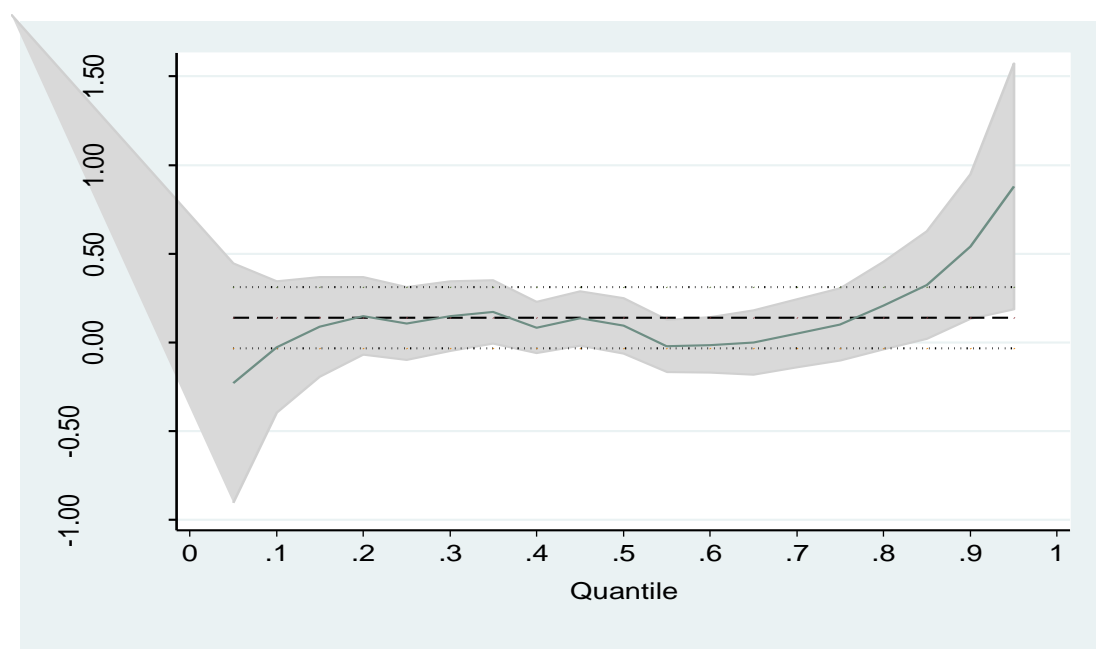


Figure 7.2: Estimates of the linear OLS and QR models.

This figure shows estimates of the OLS and the QR of the effect of shift in sentiment (bullishness) on stock returns. The dashed line represents the OLS estimate along with its 95% confidence interval (dotted lines). The x-axis represents the coefficient estimate for the change in bullishness whereas the y-axis represents the quantiles distributions of return ($\tau = 0.05, 0.1, 0.25, \dots, 0.75, 0.9, 0.95$). The results of the estimated $\beta(\tau)$ parameters for quantiles 0.05 to .95 of the QR model in Eq. (9) is depicted by the green line along with their 95% confidence intervals (shaded area).

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

As can be seen from Figure 7.2, the estimated coefficients of $\beta(\tau)$ vary with the quantiles. They are negative (positive) at low (high) quantiles, although they exhibit statistical significance at higher quantiles only. In a broad sense, the results suggest that the estimated coefficients of sentiment exert statistical significance at high quantiles, while such significance tends to diminish at low (i.e., 0.05, 0.1 and 0.25) and medium (i.e., 0.4, 0.5 and 0.6) quantiles (see Table 7.5).

Moreover, regardless of the statistical significance, the magnitude of the estimated coefficient $\beta(\tau)$ generally increases (in either sign) as τ moves out from the medium quantile towards the lower and upper quantiles. This implies that as it moves away from the 0.5 percentile towards estimates in the tails of the return distribution, the impact of sentiment changes markedly. Thus, sentiment exerts different effects on the two sides of the return distribution, with such effects becoming stronger at the very extreme quantiles (0.05 and 0.95). The size of these coefficients in absolute value is larger at the high quantiles (i.e., $\beta(0.95) = +0.880$) compared to the lower ones (i.e., $\beta(0.05) = -0.230$).

These results do not corroborate those of the OLS estimate, which shows a positive but insignificant effect of sentiment (see Table 7.5). The insignificant effect indicates that sentiment does not contain information in explaining the mean of asset returns. Brown and Cliff (2004) and Solt and Statman (1988) also failed to provide evidence of the significant effect of sentiment on returns. The positive but insignificant estimates of the causal effect of the OLS suggest no causality in mean between sentiments and returns. This is clear evidence that the OLS estimates do not tell the whole story and convey little information about the existent relationship. The QR, by contrast, shows that the value of the estimated coefficient of shift in sentiment varies over the conditional quantiles of the return distribution (see Table 7.5).

Since the estimated coefficients of $\beta(\tau)$ are found to be positive and statistically significant at higher quantiles, these means that an asset's return increases as investors become more bullish. Similarly, Brown and Cliff (2005) find that stock market developments are positively correlated with investor sentiment. This result implies that the "hold more" effect tends to dominate the "price pressure" effect. This means that when noise traders become more bullish about a particular security, their optimism induces traders to hold more of the risky assets than the fundamentals suggest, thereby increasing their expected returns relative to the market risk bearing.

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

Similarly, as noise traders become more bearish they tend to be pessimistic and tend to hold fewer risky assets, resulting in them lowering their expected returns and thereby selling off securities. The finding is also consistent with the investor sentiment model developed by DSSW (1990), who claim that irrational noise traders can cause the asset price to deviate from its fundamental value temporarily, after which it will revert to the mean as a result of adverse trading by arbitrageurs against them. The finding of this research strongly supports their theory of the impact of noise traders in which irrational noise traders with erroneous stochastic beliefs have the power to change the stock market prices in the capital markets and earn higher/lower expected returns.

Following Buchinsky (1998), this research also performs a symmetric quantiles test to examine whether the sentiment-return relations at the τ -th and $(1-\tau)$ -th quantiles are symmetric about the median, i.e., $\beta(\tau)+\beta(1-\tau)=2\beta(0.5)$. That is, the following equation is tested to determine whether is sufficiently close to zero.

$$\hat{\lambda}_T(\tau) = \widehat{\beta}_T(\tau) + \widehat{\beta}_T(1-\tau) - 2\widehat{\beta}_T(0.5) \quad (7.14)$$

The restriction in Eq. (7.14) is set for the pair of τ as of (0.05, 0.95), (0.1, 0.9), . . . , (0.45, 0.55). Note that the standard error of $\hat{\lambda}_T(\tau)$ is obtained using the bootstrap method. The results of testing symmetry of pairs of quantiles are presented in Table 7.6.

The null hypothesis of symmetric causal effects cannot be rejected at the 5% level for all pairs of τ except for $\tau = (0.05, 0.95)$ (see Table 7.6). Thus, the effects of the shift in sentiment are almost all symmetric around the median. The symmetry of these quantiles also provides an explanation of the insignificant OLS estimate (Table 7.6), since the positive and negative effects at the corresponding upper and lower quantiles tend to explain the non-causality effect in the mean as discussed earlier.

The findings broadly indicate a noticeable inverted S-shaped pattern of the sentiment coefficient estimates across quantiles of the return distribution: shift in sentiment exhibits significant positive effects on returns at higher quantiles but negative and insignificant effects on average at lower quantiles. The results imply that the impact of investor sentiment on stock returns depends on the state of the market

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

(i.e., low vs. high quantiles of returns). In a broad sense, these findings are in line with those of Baur et al. (2012), who find an inverted S-shaped pattern of the autoregressive coefficient estimates for daily and monthly returns. Their findings revealed that the autoregressive parameter markedly changes across the various quantiles of the conditional return distribution: lower quantiles exhibit on average negative dependence on past returns while upper quantiles are marked by positive dependence.

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

Table 7.5: Estimated coefficients of the linear OLS and QR models

The estimated OLS and QR models are respectively specified as $R_{it} = \alpha + \beta \Delta \bar{B}_{it}^{**} + \gamma_1 MKT_t + \gamma_2 NWK_t + \varepsilon_{it}$, and $R_{it} = \alpha(\tau) + \beta(\tau) \Delta \bar{B}_{it}^{**} + \gamma_1(\tau) MKT_t + \gamma_2(\tau) NWK_t + \varepsilon_{it}(\tau)$, where R_{it} is the returns and $\Delta \bar{B}_{it}^{**}$ is the shift in bullishness index (the proxy of investor sentiment). Total number of observations is 7,530, with a sample period from April 4th 2012 to April 5th 2013.

Level	OLS	Low			Medium			High		
		τ	0.05	0.10	0.25	0.40	0.50	0.60	0.75	0.90
α	0.002 (0.012)	-1.408*** (0.037)	-1.008*** (0.020)	-0.465*** (0.014)	-0.152*** (0.009)	0.012 (0.009)	0.186*** (0.009)	0.488*** (0.014)	1.016*** (0.021)	1.381*** (0.033)
β	0.139 (0.088)	-0.230 (0.304)	-0.026 (0.248)	0.106 (0.167)	0.084 (0.115)	0.094 (0.115)	-0.015 (0.094)	0.102 (0.106)	0.542*** (0.159)	0.880*** (0.233)
γ_1	0.998*** (0.014)	1.055*** (0.029)	1.019*** (0.020)	0.989*** (0.017)	0.976*** (0.015)	0.970*** (0.015)	0.970*** (0.016)	0.959*** (0.021)	1.007*** (0.022)	1.048*** (0.027)
γ_2	0.013 (0.025)	0.066 (0.069)	0.090** (0.044)	0.024 (0.028)	0.013 (0.024)	0.004 (0.025)	-0.029 (0.021)	-0.019 (0.026)	-0.014 (0.044)	-0.014 (0.064)
R^2	0.393									
Pseudo- R^2		0.212	0.228	0.242	0.252	0.257	0.258	0.253	0.241	0.225

Notes: *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. Standard errors are represented in parentheses.

Table 7.6: Testing symmetry of quantile causal effects of the linear model

τ Pair	(0.05, 0.95)	(0.10, 0.90)	(0.15, 0.85)	(0.20, 0.80)	(0.25, 0.75)	(0.30, 0.70)	(0.35, 0.65)	(0.40, 0.60)	(0.45, 0.55)
Restriction Coefficients (β)	0.462* (0.277)	0.328 (0.256)	0.224 (0.183)	0.168 (0.167)	0.020 (0.159)	0.011 (0.134)	-0.017 (0.113)	-0.118 (0.093)	-0.072 (0.066)

Notes: Each value is a restriction coefficient for the model $\beta(\tau) + \beta(1 - \tau) = 2\beta(0.5)$ that tests the hypothesis that the estimated parameters in selected quantiles (1- τ) are symmetric around the median and are different from those in the corresponding (1- τ) quantile. Standard errors are in parentheses. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

- **Nonlinear models results**

This section provides estimates of the asymmetric effects of changes in investor sentiment on stock returns. As outlined earlier, the model, given in Eq. (7.13), is estimated using 9 quantiles, which are divided, into three levels: low, medium and high quantiles. Figure 7.3 plots the estimated coefficients of $\beta_1(\tau)$ and $\beta_2(\tau)$ which represent the shift in bullish and bearish sentiments, respectively. The corresponding numerical results of the estimated coefficients of the QR along with those of the OLS are reported in Table 7.7.

As shown from Table 7.7, the results of the OLS suggest the existence of insignificant positive effects of both bullish and bearish shift in sentiment on stock returns, whereas those of the QR appear differently since they suggest the existence of asymmetric effects at some conditional quantiles. Specifically, the results indicate that there is a significant negative (positive) impact of bullish sentiment on returns at high (low) quantiles, but not at medium ones. The impact of bearish sentiment, in contrast, is positive (negative) at low (high) quantiles, being highly significant at lower quantiles (i.e $\tau= 0.05$ and 0.1), while showing a broadly significant impact at a very high quantiles (only at $\tau= 0.95$).

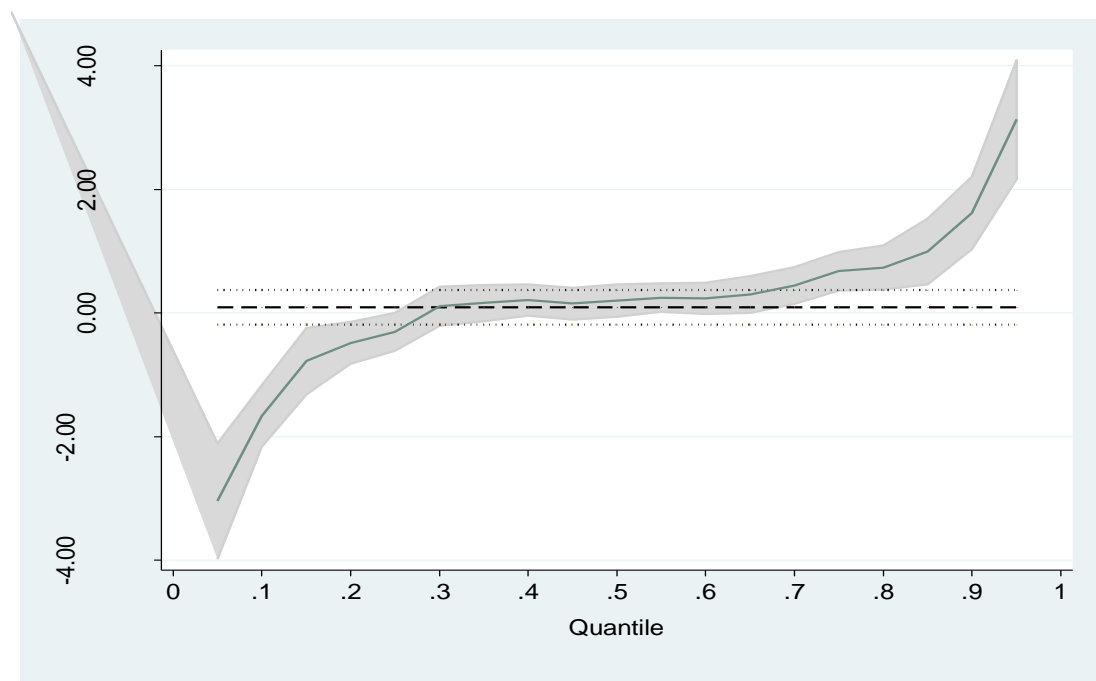
Figure 7.3 further confirms that the estimated coefficients of $\beta_1(\tau)$ and $\beta_2(\tau)$ vary with the quantiles and exhibit opposite and asymmetric behavioural patterns at the two sided of quantile distributions of returns. The graphical analysis suggests that the estimated coefficients of $\beta_1(\tau)$ (for bullish sentiment) exhibit an interesting pattern as parameters estimate exerts opposite and heterogeneous effects on the two sides of the return distribution and such an effect becomes stronger at more extreme quantiles. These results propose that the estimated coefficients of $\beta_1(\tau)$ exhibit a similar pattern to the estimated coefficient of $\beta(\tau)$ and imply a remarkable inverted S-shaped pattern of the bullish sentiment coefficient estimates across quantiles of the return distribution: bullish shift in sentiment exhibits significant negative (positive) effects on returns at lower (higher) quantiles. The results imply that the impact of bullish shift in sentiment on stock returns varies according to the market conditions (i.e., growth vs. recessions). The estimated coefficients of $\beta_2(\tau)$ (for bearish sentiment), in contrast, reveal that most of the significant evidence exists in the low quantiles but little evidence of significance indicated at high ones (only at $\tau = 0.95$). These estimated coefficients are in general positive in the low quantiles, whereas such

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

parameters tend to be negative in the high quantiles. The graphical representations of the relationship between bearish sentiment and return (by putting bearish sentiment on vertical axis and return on the horizontal axis as shown in Figure 7.3 (B)) exhibit S shape pattern across return quantiles. This relationship however is quite opposite to the similar relationship found in bullish sentiments and returns where the later shows an inverted S shape pattern across returns quantiles.

It follows that the evidence is broadly significant in the lower and higher quantiles; hence, return effects of sentiment are at play in extreme stock market conditions, i.e., sharp increases and declines. This result accords with that of Jansen and Tsai (2010), who confirm that the asymmetric effect of monetary policy surprises on stock returns is related to the bullish and bearish markets.

Panel A: OLS and QR estimates of the effect of bullish shift in sentiment on stock returns



Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

Panel B: OLS and QR estimates of the effect of bearish shift in sentiment on stock returns

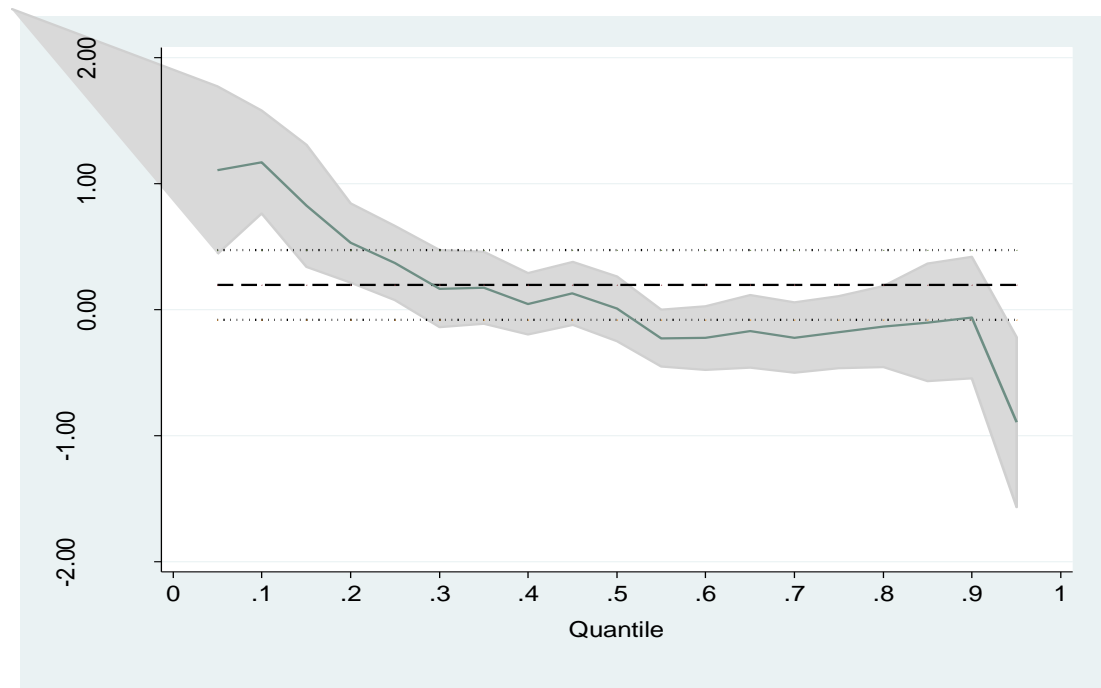


Figure 7.3: Estimates of the nonlinear (asymmetric) OLS and QR models.

This figure shows estimates of the OLS and the QR of the effect of bullish and bearish shifts in sentiment on stock returns in panels A and B, respectively. The dashed lines represent the OLS estimates along with their 95% confidence intervals (dotted lines). In both panels A and B, the x-axis represents the coefficient estimates for the change in bullish and bearish sentiments whereas the y-axis represents the quantiles distributions of return ($\tau = 0.05, 0.1, 0.25, \dots, 0.75, 0.9, 0.95$). The results of the estimated $\beta_1(\tau)$ and $\beta_2(\tau)$ parameters in panels A and B respectively, for QR model in Eq. (10) is depicted by the green line along with their 95% confidence intervals (shaded area).

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

Table 7.7: Estimated coefficients of the non-linear (asymmetric) OLS and QR models

The estimated OLS and QR models are respectively specified as $R_{it} = \alpha + \beta_1 \Delta \bar{B}_{it}^{**} D_{it} + \beta_2 \Delta \bar{B}_{it}^{**} (1 - D_{it}) + \gamma_1 MKT_t + \gamma_2 NWK_t + \varepsilon_{it}$, and $R_{it} = \alpha + \beta_1 \Delta \bar{B}_{it}^{**} D_{it} + \beta_2 \Delta \bar{B}_{it}^{**} (1 - D_{it}) + \gamma_1(\tau) MKT_t + \gamma_2(\tau) NWK_t + \varepsilon_{it}$, where R_{it} is the returns and $\Delta \bar{B}_{it}^{**}$ is the shift in bullishness index (the proxy of investor sentiment). D_{it} and $(1 - D_{it})$ are dummy variables used to capture the direction of changes (positive and negative) towards more bullish and bearish sentiments. Total number of observations is 7,530, with a sample period from April 4th 2012 to April 5th 2013.

Level	OLS	Low			Medium			High		
		τ	0.05	0.1	0.25	0.4	0.5	0.6	0.75	0.9
α	0.005 (0.015)	-1.250*** (0.042)	-0.917*** (0.025)	-0.443*** (0.020)	-0.157*** (0.015)	0.007 (0.015)	0.169*** (0.013)	0.458*** (0.015)	0.948*** (0.024)	1.257*** (0.035)
β_1	0.106 (0.146)	-3.043*** (1.123)	-1.667*** (0.403)	-0.305 (0.294)	0.212 (0.186)	0.205 (0.175)	0.240* (0.136)	0.677*** (0.181)	1.621*** (0.475)	3.133*** (0.758)
β_2	0.187 (0.145)	1.107** (0.489)	1.172*** (0.248)	0.371* (0.226)	0.048 (0.153)	0.008 (0.218)	-0.224* (0.111)	-0.177 (0.130)	-0.062 (0.276)	-0.891* (0.553)
γ_1	0.998*** (0.014)	1.065*** (0.033)	1.016*** (0.020)	0.988*** (0.017)	0.975*** (0.015)	0.968*** (0.015)	0.969*** (0.014)	0.958*** (0.019)	0.996*** (0.021)	1.030*** (0.031)
γ_2	0.013 (0.025)	0.074 (0.068)	0.089** (0.036)	0.014 (0.028)	0.013 (0.019)	0.003 (0.020)	-0.026 (0.019)	-0.021 (0.029)	-0.021 (0.050)	-0.035 (0.057)
R^2	0.393									
Pseudo- R^2		0.224	0.235	0.242	0.252	0.257	0.259	0.254	0.245	0.234

Notes: *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. Standard errors are represented in parentheses.

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

Moreover, the findings observed at low quantiles can be interpreted in the context of the DSSW (1990) model of noise traders. The model states that the change in investor sentiment as either bullish or bearish is closely related to the direction of the shift in asset returns. In particular, investors' overreactions may provide better explanations of these results of the negative (positive) impact of the bullish (bearish) shift in sentiment on stock returns. For example, if noise traders overreact to good news by becoming bullish towards a particular security, they are more likely to drive its return up. Therefore, an arbitrageur who takes an immediate action of short selling of those assets is more likely to experience low returns or even a loss at the time of selling. There are two justified reasons for this possible loss. First, noise traders may become even more bullish if additional new information of a positive nature suddenly arrives on the market, which may cause returns to rise even more. Hence, arbitrageurs' actions of immediate liquidation will cause a loss as the prices rise above those of initial short selling. Second, arbitrageurs may also bear the additional risk of a subsequent loss when they have to buy back the assets at a higher price in the future.

Another possible explanation for the significant negative effect of bullish sentiment on return at low quantiles is that bullish investors with high levels of optimism will tend to overvalue stocks by driving up prices from their fundamental levels, which consequently lowers the subsequent returns. Brown and Cliff (2005) also find that optimism is associated with overvaluation and low subsequent returns. This finding can also be explained in the context of the well-known overconfidence theory in the empirical finance literature (e.g., Miller, 1977; Odean 1998b, 1999; Gervais and Odean, 2001). The price optimism model suggested by Miller (1997) implies that an investor with overconfidence will overvalue stocks in the capital markets, causing the market prices of the stock to be relatively higher than its intrinsic value, thereby inducing investors with optimistic beliefs to trade even more, which, in turn, lowers expected returns.

On the other hand, noise traders may also overreact to bad news by becoming pessimistic or bearish towards a particular asset, thereby bringing the returns down. An arbitrageur selling this asset should recognise that noise traders may become even more bearish and drive the returns down further. That is, if arbitrageurs decide on immediate liquidation of their position in the market, they may reduce their risk exposure to such a loss. But if arbitrageurs postpone the liquidation action, noise

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

traders may become even more bearish, which may cause arbitrageurs to suffer a greater loss. In this case, the reduction in risk exposure can be interpreted as a “gain” which explains the positive returns reliance on the bearish shift in sentiments indicated by positive estimates of $\beta_2(\tau)$ at lower quantiles.

At higher quantiles, this research finds that bullish (bearish) shift in sentiments exhibits significant positive (negative) impact on returns. The positive effect of the bullish shift in sentiments on returns could be explained by the fact that when noise traders hold an optimistic belief about a traded security, they tend to hold or buy more of risky assets than fundamentals would be. Holding relatively more of risky assets would result in an increase in their expected returns relative to the market risk bearing. These results might therefore be explained by the risk return-trade-off concept. Where in the capital market, the more risk noise traders can afford through holding more risky assets as their sentiment becomes more bullish, the higher the expected return they may get.

Instead the research finds a decline in returns as sentiment becomes more bearish. This confirms the noise trader theory of DSSW (1990) in which the likelihood of holding risky assets decreases when noise traders are more pessimistic. This negative return effect caused by sentiment induces the sale of securities. As with bullish sentiments, the risk return trade off concepts implies that the less risky assets noise traders choose to hold as a result of their bearish sentiment, the less return they may expect from the sale of securities.

The discussion so far indicates that investor sentiment affects stock returns; hence stock markets are not efficient. Our findings reveal that the dynamic behaviour of asset returns varies according to the magnitude of dispersion in returns caused by the shift in investor sentiment as either bullish or bearish. It follows that the behavioural non-linear dynamics of sentiment on different quantiles of the return distribution are indeed at play.

When comparing the estimated coefficients of bullish and bearish sentiments in Eq. (7.13) separately with the estimated coefficient of sentiment in the quantile regression model, given in Eq. (7.12), there are in fact some similarities and differences in the behaviour of these estimates in relation to the returns. For example, the estimated coefficients of $\beta_1(\tau)$ (bullish sentiment) exhibit a similar pattern to

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

$\beta(\tau)$ by showing a negative (positive) influence on returns at lower (higher) quantiles, hence suggesting an inverted S-shaped pattern in affecting returns. However, they are different in terms of the statistical significance in affecting returns: $\beta_1(\tau)$ exert a statistically significant impact on returns both at lower and higher quantiles, whereas $\beta_2(\tau)$ show a significant influence only at higher quantiles. As far as the estimated coefficients of $\beta_2(\tau)$ (bearish sentiment) are concerned, the coefficient estimate of $\beta_2(\tau)$ are different than those of $\beta_1(\tau)$ in that they exhibit an opposite pattern on return, suggesting positive (negative) effects at low (high) quantiles showing a statistical significance only at lower quantiles.

Table 7.8 summarises the results of testing the symmetry of the pairwise quantile causal effects for bullish and bearish sentiments. The results indicate that, for both behavioural sentiments, these effects are all symmetric around the median, except for the second pairs in the bearish sentiment (i.e., $\tau = (0.1, 0.9)$). Indeed, the parameter estimates in the 0.1 quantile are significantly different from those of the corresponding 0.9 quantile, thereby causing such quantile to be non-symmetric about the median

Table 7.8: Testing symmetry of quantile causal effects of the non-linear (asymmetric) model

τ Pair	(0.05, 0.95)	(0.10, 0.90)	(0.15, 0.85)	(0.20, 0.80)	(0.25, 0.75)	(0.30, 0.70)	(0.35, 0.65)	(0.40, 0.60)	(0.45, 0.55)
Restriction Coefficients (β_1)	-0.319 (1.308)	-0.455 (0.597)	-0.188 (0.576)	-0.152 (0.290)	-0.038 (0.282)	0.150 (0.244)	0.051 (0.208)	0.043 (0.163)	-0.005 (0.138)
Restriction Coefficients (β_2)	0.199 (0.702)	1.094** (0.398)	0.709 (0.431)	0.381 (0.353)	0.178 (0.346)	-0.070 (0.310)	-0.012 (0.279)	-0.193 (0.285)	-0.112 (0.245)

Notes: Each value is a restriction coefficient for the joint model of $\beta_1(\tau) + \beta_1(1 - \tau) = 2\beta_1(0.5)$ and $\beta_2(\tau) + \beta_2(1 - \tau) = 2\beta_2(0.5)$ that tests the hypothesis that the estimated parameters in selected quantiles τ are symmetric around the median and are different from those in the corresponding $(1 - \tau)$ quantile. Standard errors are in parentheses. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively

The empirical results discussed in this section are summarised as follow. First, the OLS estimate shows that the shift in investor sentiment contains no information that explains stock returns in the capital markets. The QR model adopted in this study, in contrast, provides evidence that the impact of sentiment on returns varies across the quantiles of the return distribution. In particular, a shift in sentiment exerts a negative (positive) effect at low (high) quantiles. Second, the research confirms the existence of asymmetric effects of sentiment on stock returns. The bullish and bearish shift in sentiments exerts an opposite and heterogenous effect on the two tails of the return

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

distribution whereby significant evidence is found for extreme stock market conditions, in particular. These results are broadly in line with those of Brown and Cliff (2005), Wang (2001) and Hong et al. (2000), who reveal that the effect of the sentiments of noise traders on stock prices can be asymmetric.

7.5 Investors' divergence of opinion and Trading Volume

Trading volume is one of the most important measures of the financial market and it has received attention in most of the empirical finance literature. One of the most important aspects of trading volume that has been widely discussed in the literature since 1977 is whether disagreement produces trading (Hirshleifer, 1977; Harris and Raviv, 1993). The role of disagreement in trading volume is harder to predict. The implications of divergence of opinion in the stock market have been addressed by Miller (1977). He theorised that, with the existence of short-sell constraints and the disagreement among investors, the prices will reflect only the valuation of the most optimistic investors and not the pessimistic ones. In the market, if the short-sell constraint binds, investors with high valuations will not short-sell the stock; they will either sell the shares or stay out of the market if they agree with the market price. Miller's model suggests that a greater divergence of opinion induces trading and leads to higher market prices, compared to the real value of the stock, and lower future returns.

The central concern of this section, therefore, is to investigate the relationship between investors' disagreement and trading volume. The principal objective in this section is to investigate the role of disagreement in online investors' opinions in predicting trading volume in the financial market. In this section, both the linear and non-linear effects of disagreement on trading volumes are considered. That is, the asymmetric effect of the divergence of investors' opinions on trading volume is estimated in two different regimes/patterns of return: bull and bear markets. The aim is to capture the asymmetry in the predictive power of investors' disagreement in trading volume in these two regimes individually.

7.5.1 Empirical Methodology and model specifications

- **The linear model**

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

The model that explores the volume-disagreement relations based on contemporaneous effect is described in the following equation:

$$TV_{it} = \alpha_1 + \phi_1 TV_{it-1} + \lambda_1 A_{it} + \beta_1 NWK_t + \beta_2 MKT_t + \varepsilon_{it} \quad (7.15)$$

where TV_{it} is the trading volume of stock i on day t , A_{it} is the level of agreement of stock i at time t , MKT is the market return of INDU index (DJIA index return) added in the regressions as a control variable to control for the overall market-wide effects, and NWK is a dummy variable for the first day of the trading week to control for potential return anomaly effect.⁴³ Panel regressions with company fixed effects are estimated using standard ordinary least squares (OLS) technique. The OLS estimates of the coefficients λ_1 in Eq. 7.15 are the primary focus of this regression model. These coefficients describe the dependency of trading volume on the level of agreement using StockTwits messages.

- **The asymmetric model**

Following the intuition of Varian (1985), Harris and Raviv (1993), and Shalen (1993), who provided the first empirical evidence of the role of variation in investors' opinions regarding the interpretation of arriving news, this study investigates the predictability pattern of investors' disagreement on trading volumes across two different states of the market: the bull and bear markets. It is believed that the magnitudes of the effects of disagreement on trading volume are likely to be asymmetric in the two stated regimes.

To test the asymmetrical effect of the impact of disagreement on trading volumes in the two different states of the market, two indicator variables are created and labelled I_{it}^{bull} and I_{it}^{bear} for the bull and bear markets respectively. These two indicator variables were previously defined in Eq.7.7 and 7.8 in Section 7.3 of this chapter. (For more details please refer to Section 7.3).

In order to investigate the significance of the explanatory power of disagreement on trading volume in the two different states of the market, two

⁴³ Day after holiday' effect is one of the stock return anomalies where returns on stocks are found to be lower; i.e. lower on Mondays than on other days of the week (Thalar, 1987). To control for this return anomaly a dummy variable is created that takes the value of one on the day after a holiday and zero otherwise.

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

interaction terms of each indicator variable with the measure of disagreement $I^{bull} A_{i,t}$ and $I^{bear} A_{i,t}$ are created and added to the volume regression, as follows⁴⁴:

$$TV_{it} = \alpha_1 + \phi_1 TV_{i,t-1} + \gamma_1 I_{it}^{bull} \cdot A_{it} + \gamma_2 I_{it}^{bear} \cdot A_{it} + \beta_1 NWK_t + \beta_2 MKT_t + \varepsilon_{it} \quad (7.16)$$

where $I_{it}^{bull} \cdot A_{it}$ and $I_{it}^{bear} \cdot A_{it}$ are the two interaction terms of the disagreement measures in the bull and bear markets, respectively. The coefficients estimates of the interaction terms measured by γ_1 and γ_2 are the main concern in the model shown in Eq. (7.16).

7.5.2 Volume portfolio strategies based on disagreement

One of the major advantages of the volume-disagreement model used here is that it provides inclusive understanding of not only the direct impact of disagreement on trading volume but also of the effect of such impact on the behaviour of asset prices in the capital market. To justify the support for this model, a portfolio strategy is constructed in which stocks are assigned to different portfolios according to some predetermined characteristics, such as traded shares and or the level of disagreement among messages. The aim is to confirm the comprehensibility of this study's compelling model that allows an investigation of the simultaneous effect of the impact of disagreement on behaviour of stock prices and trading volume.

To ensure that the results of this research portfolio formation are not driven by small, illiquid stocks or by bid-ask bounce, this study follows the approach of Jegadeesh and Titman (2001), where stocks with share prices lower than five dollars are removed from the sorting process. The methodology employed here follows the two-way sort as conducted by Diether et al. (2002), who use more than a one-way sort (both the double and triple sort) in creating the portfolio. However, this study only uses a double sort based on the volume of traded shares and disagreement. This process is performed as follows. First, in each month the sample stocks are assigned to five quintiles based on the volumes of traded shares as of the previous month. V1

⁴⁴ The estimated result of this model shows that the model specification suffers from serial correlation. To correct the model for serial correlation, one-day lag of trading volume is therefore added to the volume regression.

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

comprises the stocks with lowest volumes of trade, while V5 contains the stocks with the highest trading volumes. Then, each of these portfolio groups is further sorted into five quintiles based on the disagreement measure as of the previous month. A1 comprises stocks with low agreement level (high disagreement) while A5 contains stocks with a high level of agreement (low disagreement). The two-way sort results in the assigning of stocks to 25 portfolios. The stocks are held in the portfolios for the entire trading month and are then re-sorted at the beginning of the next trading month based on a new level of trading volumes. The monthly portfolio returns are calculated as equally weighted average returns of all stocks in the portfolio. This two-way portfolio sorting strategy is repeated at the end of each month over one year of the study sample period.

The last column of Table 7.9 shows that this type of portfolio sorting produces a strong positive relation ($D1 \rightarrow D3$) between average returns and disagreement for the stocks in the low disagreement portfolios, while a negative relation ($D3 \rightarrow D5$) between returns and disagreement is likely to be found in stocks in the high disagreement portfolios. This result supports the model of Miller (1977), who argues that, the higher the disagreement about the value of a stock, the higher the market value relative to the intrinsic value of the stock, and the lower its expected returns. The average monthly portfolio return differential between low-high disagreement (D1-D5) portfolios declines as the average trading volume increases. While lower returns are presented for each volume quintile, the order of magnitude of the difference appears to be a decreasing function of volume. In particular, the returns of D1 (low)-D5 (high) disagreement strategy range from 0.082 for low-volume stocks to -0.105 for high-volume stocks. This result signifies that the influence of the disagreement on stock returns is more pronounced in the highly traded stocks. The t-statistics of the return differential between the low- and high-disagreement stocks in the last row of Table 7.9 are positive (negative) and highly significant for the extreme low (medium and extreme high) volume stocks, whereas this becomes insignificant for stocks in the second and fourth volume quintiles.

Closer inspection of Table 7.9 reveals that low disagreement portfolios of both small and heavily traded stocks result in a decrease in stock returns. In other words, the disagreement-return relation in the low disagreement portfolios does not show a significantly different effect on average monthly returns between the small(V1V2)

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

and heavily traded shares (V4 → V5). However, the medium traded share (V3 → V4) of the low disagreement portfolios results in an increase in stock returns. On the other hand, the high disagreement portfolios show a significantly different effect on returns for the small (V1 → V2) and medium (V3 → V4) than for (highly) traded shares (V4 → V5) where an increase (decrease) in average monthly returns pronounced for small, medium and (heavily) traded stocks respectively.

Table 7.9: Mean Portfolio returns by Trading Volume and Disagreement

This table shows the two-way portfolio sorting based on trading volume and the level of disagreement. Each month stocks are sorted in five quintiles based on the trading volumes at the end of the previous month. Stocks in each volume quintiles are then resorted into additional five quintiles based on the level of disagreement of each month. All Stocks will be assigning into 25 portfolio groups, which are then held for one month. The monthly weighted average returns are calculated for each portfolio group. The Sample period considered is from 3rd of April 2012 to 5th of April 2013 of DJIA stocks.

Mean Return						
Volume Quintiles						
	Small			Large		
Disagreement Quintiles	V1	V2	V3	V4	V5	All Stocks
D1(Low)	0.105	0.082	-0.090	0.138	-0.097	0.026
D2	0.169	0.027	0.038	0.074	-0.042	0.055
D3	0.169	0.119	0.018	0.208	0.024	0.107
D4	0.057	0.131	0.015	0.034	-0.047	0.032
D5(High)	0.023	0.075	-0.004	0.187	0.008	0.055
D1 – D5	0.082**	0.007	-0.086***	-0.049	-0.105***	0.029**
(t-Statistics)	(3.124)	(0.427)	(-4.299)	(-1.662)	(-5.450)	(-2.216)

The t-statistics in parentheses test whether the mean of differences is equal to zero. (**) (***) indicates significance at the 1% and 5% levels, respectively.

In summary, the results of the volume-disagreement portfolio strategy are successfully linked to the relationships between the three variables studied: trading volumes, returns and disagreement. These findings provide great support for the heterogeneous-agent model by successfully showing that volume-disagreement relations make it possible to test for the impact of disagreement not only on trading volume but on stock price behaviour (i.e. returns) simultaneously. This suggests that the extensive literature on volume and stock returns might also be explained by and/or linked to the literature on the disagreement. The investigations in this study concluded

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

that high disagreement results in higher (lower) returns for the lightly (heavily) traded stocks respectively and that occasional large divergences in prices (returns) may be part of the same phenomenon. This suggests that a consistent explanation for one result may offer an explanation for the other.

7.5.3 Empirical Findings and Regression Result

- **The linear model**

The linear model estimates the contemporaneous correlation of the impact of disagreement on trading volume given in Eq. (7.15). Table 7.10 presents the regression estimate output of the panel OLS regression estimated using the company fixed effect. The parameter estimate of interest in this equation is λ_1 , which measures both the direction and the magnitude effect of the possible influence of disagreement (the proxy for divergence of opinion) on trading volume.

The results presented in Table 7.10 suggest that the model specification of volume- disagreements in Eq. (7.15) suffers from serial autocorrelation in the data series. Therefore, the model is modified whereby a lagged trading volume is included in the volume regression to assist in removing the serial correlations (Dickey and Fuller, 1979; Balvers et al., 2000).⁴⁵ The impact of the disagreement on trading volume is challenging to predict, yet this study's results are in line with those of Antweiler and Frank (2004b) and Sprenger et al. (2014), who found negative correlations between agreement and trading volume. This suggests that greater agreement between the buy and sell messages in a period is associated with fewer trades during that period. In other words, this negative relationship implies that the high disagreement among messages on a given stock triggered increases in the traded shares of that particular stock. This means that when the level of disagreement among messages increased by one unit of standard deviation (standard deviation of the sell and buy messages), this would cause 5.2% increases in the level of trading volumes. The negative estimated coefficient of ($\lambda_1 = -0.052$, $p\text{-value} < 0.001$) generally supports the first hypothesis of this study, which states that disagreement is positively correlated with trading volume. Therefore, the result provides support for what has

⁴⁵ This study experiences serial autocorrelations in all volume-disagreement models; therefore, the lagged trading volume is added to all of the corresponding regression equations. It is found that adding only one-day lag trading volume solves the problem of the serial correlation in this date series.

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

been previously deduced theoretically (e.g. Harris and Raviv, 1993; Karpoff, 1986; Kim and Verrecchia, 1991) that disagreement among traders resulted in an increase in trading volume. This signifies the explanatory power of disagreement in explaining the trading volume.

Table 7.10: The linear Regression Model (volume-disagreement)

This table reports the results of the trading volume model of panel data of DJIA index over the period April 3, 2012 to April 5, 2013. Using the agreement Index extracted from StockTwits postings (the disagreement measure is used as a proxy for divergence in opinion among investors). This Model incorporates the contemporaneous of the impact of disagreement among investors measured by A_t on trading volume. Dummy variables are added to capture first day of the trading week effect to control for Monday return anomaly. Market returns of DJIA index are added as a control variable to control for market-wide effect. The equation is estimated based on OLS panel regression with company fixed effect. The reported coefficients are for the following regression:

$$TV_{it} = \alpha_1 + \phi_1 TV_{it-1} + \lambda_1 A_{it} + \beta_1 NWK_t + \beta_2 MKT_t + \varepsilon_{it}$$

Volume Regressions	Contemporaneous regression at time (t)	
α_1	9.752***	(0.161)
ϕ_1	0.335***	(0.011)
λ_1	-0.052***	(0.019)
β_1	-0.153***	(0.010)
β_2	-0.023***	(0.006)
R^2	0.844	
Durbin-Watson statistics	2.085	
N- observation	7,530	

Note (*), (**), and (***) denote significance levels at the 10%, 5%, and 1%, respectively. Standard error is shown in parenthesis.

- **The nonlinear model**

One of the most challenging relationships to anticipate is the role of disagreement in explaining trading volume in the stock market. The predictive model in the previous section provides results, which corroborate the findings of a great deal of the previous work in this field. The findings of this study support the traditional hypothesis of Harris and Raviv (1993) that disagreement induces trading, indicated by the negative coefficients of the agreement index in the trading volume regression both

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

in the contemporaneous and time sequencing relationships⁴⁶. However, the aim in this section is to investigate whether this study might provide evidence of a possible asymmetric effect of the impact of disagreement on trading volume in two regime classifications of the market (bull and bear markets).

The central concern from Table 7.11 is the coefficients of the interaction terms I_{it}^{bull} , A_{it} and I_{it}^{bear} , A_{it} (measured by γ_1 and γ_2) in Eq. (7.16) which measures the impact of disagreement on trading volumes in the bull and bear markets, respectively. The results found that both interaction terms' parameters exert negatively and statistically significant effects in explaining trading volume, indicated by ($\gamma_1 = -0.038$ p-value < 0.1 and $\gamma_2 = -0.066$, p-value < 0.01). This could be interpreted as following; a one unit standard deviation increase in the level of disagreement among the buy and sell messages leads to a 3.8% and 6.6% increase in the trading volumes in the bull and bear market, respectively. More interestingly, all of the disagreement coefficients' estimates exert statistically significant negative effects on trading volume in the bull and bear markets. This result implies that, in good-news and bad-news environments, high disagreement triggers intensity in volumes of traded shares in the capital market. That is, high disagreement among traders produces abnormal trading activities irrespective of the type of news.

These study findings can be explained in the context of over-confidence theory and short-sell constraints suggested in a great deal of the empirical literature (e.g. Miller, 1977; Odean 1998b, 1999; Gervais and Odean, 2001). For example, if noise traders experience a positive increase in return (as good news arrives) of a given stock in the market, they are more likely to become optimistic about that particular security, thus supporting the positive feedback trading theory (De Long et al., 1990). Meanwhile, the price optimism model suggested by Miller (1977) implies that, in a period of good news, higher disagreement among traders will cause the market price of the stock to be relatively higher than its intrinsic value, which causes investors with optimistic beliefs to express their over-confidence through higher trading. At the same time, other investors with pessimistic beliefs will be subject to short-sell constraints, leading them to sell all their shares and keeping them out of the market. As a consequence, the stock price will only reflect the opinions of optimistic investors

⁴⁶ Although the results are not reported herein, the time sequencing relationship (one period lag) between disagreements and trading volumes is also tested and the results are in line with the contemporaneous results that confirm the positive effect of disagreement on trading volumes.

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

whereas the opinions of pessimistic investors will be disregarded. Therefore, in the bull market the high disagreement among traders will result in simultaneous buying (selling) activities by bullish (bearish) investors, respectively, triggering a higher intensity of trading volume in the capital market.

In contrast, in a period of market decline the higher disagreement among bullish and bearish investors also triggers an increase in the trading volumes. Hong and Stein (1999) argue that the large differences of opinion among traders will force the bearish investors who are subject to the short-sell constraints to keep out of the market while the stock prices can only reflect the views of investors who are too optimistic. As a result of being at the side, the market prices will not fully incorporate the information of those constrained investors, and their information will therefore be hidden. The earnings announcement news, however, will force this hidden information out into the market, causing reversed trading roles between the previously more bullish investors and previously bearish investors. The inverted round of trade at this time will cause investors with prior optimistic beliefs to bond out of the market while the previously bearish group of investors will be the “support buyers”. At this stage, future events news (e.g. earnings announcement news) will be regarded as bad news by those previously more bullish investors who set out a market price above its true value, thereby experiencing negative market returns. Accordingly, as the hidden information is forced to come out during market declines, previously bearish investors will be active buyers whereas originally bullish investors are forced to short-sell and bail out of the market. This price behaviour explains the profitability of contrarian strategies whereby past losers (previously bearish investors) outperform past winners (previously bullish investors), as suggested by De Bondt and Thaler (1985) and Jegadeesh and Titman (2001). Therefore, these results imply that, in the bear market, greater disagreement among traders will result in concurrent buying (selling) activities by bearish (bullish) investors, respectively, triggering a higher intensity of trading volume in the capital market.

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

Table 7.11: The non-linear model (asymmetric response of trading volume to the investors' disagreement in the bull and bear market)

This table reports the results of the asymmetric effect of disagreement on trading volumes in two different market regimes (the bull and bear market) using panel data of 30 companies of DJIA index over the period April 3, 2012 to April 5, 2013. The agreement index is extracted from StockTwits postings where the disagreement among traders is used as a proxy for divergence of investors' opinions. The model that incorporates the contemporaneous effect of the impact of disagreement on trading volume is estimated. The reported coefficients are for the following regressions:

$$TV_{it} = \alpha_1 + \phi_1 TV_{i,t-1} + \gamma_1 I_{it}^{bull} \cdot A_{it} + \gamma_2 I_{it}^{bear} \cdot A_{it} + \beta_1 NWK_t + \beta_2 MKT_t + \varepsilon_{it}$$

where; $I_{it}^{bull} \cdot A_{it}$ and $I_{it}^{bear} \cdot A_{it}$ are the two interaction terms of the disagreement measures in the bull and bear markets, respectively. Market returns of DJIA index are added as a control variable to control for market-wide effect. The regressions are estimated based on OLS panel estimate with company fixed effect.

Volume Regressions	Contemporaneous regression at time (t)	
α_1	9.751***	(0.161)
ϕ_1	0.335***	(0.011)
γ_1	-0.038*	(0.025)
γ_2	-0.066***	(0.026)
β_1	-0.153***	(0.010)
β_2	-0.024***	(0.006)
R^2	0.844	
Durbin-Watson stat	2.086	
N- observation	7,530	

Note: (*), (**), and (***) denote significance levels at the 10%, 5%, and 1%, respectively. Standard error is shown in parenthesis.

It is apparent that there is symmetric response to disagreement on trading volumes in the bull and bear markets where the trading activity shows a symmetric⁴⁷ response it increases in both up and down markets⁴⁸. While the impact of disagreement on trading volume shows a similar directional effect (positive effect) in both bull and bear markets, their affected magnitudes differ asymmetrically. The magnitude of the impact of disagreement on trading volume shows a larger effect in case of the period of negative returns (bear market $\gamma_2 = -0.066$) compared to the period of positive

⁴⁷Note that symmetric response is the exactly the opposite of asymmetric effect. A nonlinear relationship is said to be symmetric if it shows a similar effect of the two states being studied (in this case the bull and bear market) in contrast to asymmetric effect that shows a different effect of these different states.

⁴⁸Note that the symmetric response of the effect of disagreement on trading volumes is being explained as having a positive effect (As shown in Table 7.11), which is exactly an inverted effect of the negative coefficients of agreement on trading volume in the bull and bear markets shown in Eq. 7.16.

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

returns (bull market $\gamma_1 = -0.038$). This suggests that volume-disagreement concerns are more important in the bear market than in the bull market. This finding supports Hong and Stein's model (1999), which predicts that negative skewness in returns tends to be most pronounced around the period of heavy trading volumes.⁴⁹ These results are also in agreement with Kling and Gao's (2008) finding that negative returns have a stronger impact on investor sentiments (and therefore disagreement) than positive returns. This is also in line with the findings on small shareholders (Sias, 1997). In general, the findings confirm the existence of symmetric effect of the impact of disagreement among investors on volumes of trade since there is a greater positive impact of disagreement on traded volumes during the period of decline in the bear market than the effect during the period of growth in the bull market.

- **Controlling for additional market factors**

While the baseline predictive model has suggested that the level of disagreement among traders contains value-relevant information in explaining trading volumes in the capital market, it is still not known whether these relationships can survive the inclusion of a more inclusive set of relevant volume determinant factors considered by Chordia et al. (2001). Other control variables following Antweiler and Frank (2004b) are also included in the volume regression in addition to the disagreement measure. The factors to be considered in the volume regression include previous changes in the stock price (stock up yesterday, stock down yesterday, stock up last 5 days and stock down last 5 days), the market index (market up yesterday, market down yesterday, market up last 5 days and market down last 5 days), and the stock and market volatility (stock 5 days' volatility, market 5 days' volatility). In line with Sprenger et al. (2014), this research also expands the list of control variables to include the change in the trading volumes in the preceding days by adding the trading volume (trading volume up yesterday, trading volume down yesterday, trading volume up last 5 days and trading volume down last 5 days). The federal funds rate (FFR), the quality spread (the difference between the yield on Moody's Baa or better corporate bond yields and the treasury rate), and the term spread (the difference between the FFR and the 10-year Treasury bill rate) are also added to the volume regression model. To capture the day-of-the-week effects in the volume regression, a

⁴⁹ High trading volumes are used as a proxy for the intensity of disagreement in many divergence of opinion models used in the empirical literature (See Varian, 1989; Harris and Raviv, 1993; Kandel and Pearson, 1995; Odean, 1999)

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

series of dummies for Monday, Tuesday, Wednesday, and Thursday is added. In addition, a dummy for days preceding or following a public holiday except when that trading day falls on a Monday or Friday was also added to the volume regression⁵⁰. The volume regression model incorporates the one-day lag of the agreement measure plus other volume determinants suggested by the empirical work of Chordia et al. (2001), Antweiler and Frank (2004b), and Sprenger et al. (2014). The behavioural specification of the model is expressed in the following equation:

$$\begin{aligned}
 TV_{it} = & \alpha_1 + \lambda_1 A_{it} + \beta_1 Stock_{it}^+ + \beta_2 Stock_{it}^- + \beta_3 MA5 Stock_{it}^+ + \beta_4 MA5 Stock_{it}^- + \\
 & \beta_5 MA5 stock\ vol_{it} + \beta_6 MKT_t^+ + \beta_7 MKT_t^- + \beta_8 MA5 MKT_t^+ + \beta_9 MA5 MKT_t^- + \\
 & \beta_{10} MA5 MKT\ vol_t + \beta_{11} TV_t^+ + \beta_{12} TV_t^- + \beta_{13} MA5 TV_{it}^+ + \beta_{14} MA5 TV_{it}^- + \beta_{15} \Delta FFR_t + \\
 & \beta_{16} Qlty\ Sprd_t + \beta_{17} Trm\ Sprd_t + \beta_{18} MON + \beta_{19} TUE + \beta_{20} WED + \beta_{21} THU + \beta_{22} Holiday + \varepsilon_{it}
 \end{aligned}
 \tag{7.17}$$

where TV_{it} = Trading volume for the i th stocks in day t ; $A_{i,t-1}$ = Agreement measure relative to the dependent variable; $Stock_{it}^+$ = Stock up yesterday = $\max \{0, \ln(P_{it-1}) - \ln(P_{it-2})\}$, daily stock return if it is positive and zero otherwise.; $Stock_{it}^-$ = Stock down yesterday = $\max \{0, \ln(P_{it-2}) - \ln(P_{it-1})\}$, daily stock return if it is negative and zero otherwise.; $MA5 Stock_{it}^+$ = Stock up last 5 days = $\max \{0, \ln(P_{it-1}) - \ln(P_{it-5})\}$, the past five trading days' daily stock return if it is positive and zero otherwise; $MA5 Stock_{it}^-$ = Stock down last 5 days = $\max \{0, \ln(P_{it-5}) - \ln(P_{it-1})\}$, the past five trading days' daily stock return if it is negative and zero otherwise; $MA5 stock\ vol_{it} =$ Stock five-day Volatility $\sum_{q=1}^5 \sum_{d \in D(t)} |\ln(P_{it-q,d}) - \ln(P_{it-q,d-1})|$; MKT_t^+ = Market up yesterday = $\max \{0, \ln(\bar{P}_{t-1}) - \ln(\bar{P}_{t-2})\}$, daily index return if it is positive and zero otherwise; MKT_t^- = Market down yesterday = $\max \{0, \ln(\bar{P}_{t-2}) - \ln(\bar{P}_{t-1})\}$, daily index return if it is negative and zero otherwise; $MA5 MKT_t^+$ = Market up last 5 days = $\max \{0, \ln(\bar{P}_{t-1}) - \ln(\bar{P}_{t-5})\}$, the past five trading days daily index return if it is positive and zero otherwise; $MA5 MKT_t^-$ = Market down last 5 days = $\max \{0, \ln(\bar{P}_{t-5}) - \ln(\bar{P}_{t-1})\}$, the past five trading days daily index return if it is negative and zero otherwise; $MA5 MKT\ vol_t =$ Market five

⁵⁰ The indicator variable for holiday takes the value of one for Wednesday, July 4th 2012 (Independence Day) Tuesday, December 25th 2012(Christmas Day) Thursday, November 22nd 2012 (Thanksgiving Day) and zero otherwise.

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

day Volatility $\sum_{i=1}^5 \sum_{d \in D(t)} |\ln(\bar{P}_{t-i,d}) - \ln(\bar{P}_{t-i,d-1})|$; TV_{it}^+ = Trading volume up yesterday = $\max \{0, \ln(TV_{it-1}) - \ln(TV_{it-2})\}$, daily stock trading volume if it is positive and zero otherwise; TV_{it}^- = Trading volume down yesterday = $\max \{0, \ln(TV_{it-2}) - \ln(TV_{it-1})\}$, daily stock trading volume if it is negative and zero otherwise; $MA5 TV_{it}^+$ = Trading volume up last 5 days = $\max \{0, \ln(TV_{it-1}) - \ln(TV_{it-5})\}$, the past five trading days' daily stock trading volume if it is positive and zero otherwise; $MA5 TV_{it}^-$ = Trading volume down last 5 days = $\max \{0, \ln(TV_{it-5}) - \ln(TV_{it-1})\}$, the past five trading days' daily stock trading volume if it is negative and zero otherwise; $FFR_t = \Delta$ Federal Fund Rate = $\{\ln(FFR_t) - \ln(FFR_{t-1})\}$; $Qlty Sprd_t =$ Quality Spread = $\Delta(BAA_t - T10_t)$; $Trm Sprd_t =$ Term Spread = $\Delta(T10_t - FFR_t)$; $MON, TUE, WED, THU =$ 1.0 Week dummies that take a value of one if the trading day is, respectively, a Monday, Tuesday, Wednesday or Thursday and zero otherwise; $Holiday =$ 1.0 if trading days are, respectively, Wednesday, July 4th 2012 (Independence Day) Tuesday, December 25th 2012 (Christmas Day) Thursday, November 22nd 2012 (Thanksgiving Day) and zero otherwise.

The volume regression model in Eq. (7.17) is estimated using OLS regression with company fixed effect. The results are reported in Table 7.12 and are mostly fairly consistent with Chordia et al. (2001) and Antweiler and Frank (2004b). Although the results show that market-wide movements are important, as noted by Chordia et al. (2001), this study provides greater support for Antweiler and Frank (2004b), in that firms' specific movements are rather more significant than market-wide movements. Unlike Antweiler and Frank (2004b), who find elevated trading at midweek, the results of this research study show that the day-of-the-week dummies are all significantly negative in the volume regression. This implies that market liquidity declines while trading activities becomes sluggish as one move on towards the end of the week. The findings of this study are in agreement with those of Sprenger et al. (2014), who show that the daily dummies reflect decreased trading, indicated by statistically significant negative coefficients of the day-of-the-week dummies in the volume regressions. The results also showed that on Monday there is a dramatic drop in the level of trading activities, indicated by the largest significant negative coefficient on Monday compared to the other days of the week. This is in line with Lakonishok and Maberly (1990) who find a reduction in the trading activities of the institutional investors at the beginning of each trading week. The

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

Quality Spread variable apparently has a very significant influence on the trading activities.

The most critical issue of concern in this type of regression analysis is whether the level of disagreement among messages remains significant in explaining the trading volumes when such large numbers of control variables are added to the volume regression. Interestingly, the investigation results in this research reveal that the level of disagreement survives the inclusions of the above discussed control variables. The findings reveal that high disagreement among traders expressed by their sentiments in StockTwits messages still exerts statistically significant positive relationships with trading volumes (negative coefficients of agreement) where high disagreement among traders triggers an increase in the level of trading activities in the capital market. Therefore, this study concluded that the level of disagreement among traders in the online stock forums contains valuable information with respect to the current trading volumes while predicting contemporaneous increases in trading volumes in the capital market.

The empirical results discussed in this section provide evidence that the level of disagreement measure (as a proxy for divergence of opinions among investors) extracted from stock-related micro-blogging forums contains relevant information that is not yet reflected in trading volumes and, hence, stock prices in the capital market. The results clearly provide significant support for the theoretical claim that high divergence in investors' opinions is associated with higher trades. They also provide evidence that large and (small) volumes of stocks with higher disagreement among traders earn significantly lower (higher) expected returns. The discussion shows that these results are consistent with the price optimism model that higher disagreement among traders will cause the market price of the stock to be relatively higher than its intrinsic value. This causes investors with optimistic beliefs to trade even more, which in turn lowers their expected returns.

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

Table 7.12: Volume Regression

This table presents the volume regression model that examine the explanatory power of the level of agreement on trading volume after controlling for large set of control variables. The independent variables include the agreement measure relative to the dependent variable. In line with Antweiler and Frank (2004b), the regressors include company-specific and market wide variables that measure the log change in daily stock returns, market returns and trading volume identically, e.g., stock up yesterday = $\max\{0, r_{t-1}, r_{t-2}\}$ where r_t represents the return on day t and stock and market 5 day volatility. Following Garman and Klass (1980), the daily volatility is estimated based upon the historical opening, closing, high, and low prices. Federal funds rate (FFR) is the US federal funds rate, the quality spread = $\ln(1+BB)-\ln(1+T10)$ where BB is the BB corporate bond yield and T10 represents the 10 year US government yield. The term spread = $\ln(1+T10)-\ln(1+FFR)$. Monday through Thursday are day of week dummies and holiday a dummy for days preceding or following a public holiday. The model is estimated using the Ordinary least square regression with company fixed effect on panel of 30 companies of DJIA in our data set. The dependent variable in this panel regression of daily company data is the log of the companies' traded shares. Note (*), (**), and (***) denote significance levels at the 10%, 5%, and 1%, respectively. Standard error is shown in parenthesis.

Independent Variables	Coefficient Estimate	Standard Error
α_1	14.4551***	(0.1597)
λ_1	-0.0345*	(0.0184)
β_1	0.0148***	(0.0057)
β_2	-0.0121**	(0.0057)
β_3	0.0300***	(0.0056)
β_4	-0.0301***	(0.0057)
β_5	0.1077***	(0.0057)
β_6	0.0018	(0.0092)
β_7	-0.0425***	(0.0094)
β_8	-0.0123	(0.0089)
β_9	0.0366***	(0.0088)
β_{10}	1.1743***	(0.2062)
β_{11}	0.0258	(0.0230)
β_{12}	0.0317	(0.0298)
β_{13}	-0.0655***	(0.0147)
β_{14}	-0.1289***	(0.0176)
β_{15}	0.1681	(0.1249)
β_{16}	-0.0010	(0.1015)
β_{17}	0.2886***	(0.1108)
β_{18}	-0.3484***	(0.0137)
β_{19}	-0.2719***	(0.0133)
β_{20}	-0.2624***	(0.0130)
β_{21}	-0.1914***	(0.0129)
β_{22}	-0.4361***	(0.0382)
R^2	0.8549	
N- Observation	7,410	
Durbin Watson	1.802	

Chapter Seven: An Empirical Investigation of The Role of Investor Sentiment in The Stock Market

7.6 Chapter Summary

This chapter provided the results and analysis of the empirical findings that test the degree to which investor sentiment affects stock return, volatility and trading volume while highlighting the role played by noise traders in determining assets pricing in the capital market, motivated by the DSSW model (1990). Studying the non-linear relationship between investor sentiments and stock return, this chapter investigated the asymmetric responses of the investor sentiments to the change in stock return in different states of the market (bull and bear markets). The aim was to examine whether investor sentiments respond differently to the good and bad news in the market, indicated by positive and negative returns in the bull and bear markets, respectively. This chapter also reported the in-depth analysis of some relationships between tweet features and stock market variables such as those of bullishness and returns and disagreement and trading volume, as these two respective relationships have created a great puzzle in the empirical literature. A quantile regression analysis coupled with the non-linear model to investigate sentiments-return relations makes a significant contribution to the existing literature. As for the volume-disagreement relationship, this study contributes by asserting that the measure of the level of disagreement extracted from an online stock micro-blogging forum represents a useful empirical proxy for the cross-sectional dispersions of investors' opinions related to underlying security values. Consistent with this assertion, the findings of this research study provide evidence that the online divergence of the traders' opinion measure contains value-relevant information that is not yet reflected in securities' trading volumes. Additionally, the analysis takes a step further to investigate the possibility of an asymmetric impact of disagreement in two different states of the economy (up/bull market and down/bear market) to investigate whether or not the relation is found to be asymmetric with respect to different market condition.

CHAPTER EIGHT: CONCLUSION

8.1 Introduction

Following the analysis of the research findings and discussions in the previous three chapters, this chapter presents the conclusions and implications of the research findings. The objective of this study was to examine the extent to which investor sentiments and other Stock Micro-blogging features (i.e. message volumes and agreement) affect stock price behavioural movements in the financial market using various statistical analyses and machine learning techniques. Motivated by the noise trader model (DSSW, 1990), this research study was able to identify the role played by investor sentiments in determining assets returns and other financial market indicators in the capital market.

This chapter is the concluding part of this thesis. It provides a summation of the research that has culminated in this thesis. The chapter begins with an overview of this research and the key findings in Section 8.2. This is followed by a discussion of the research contributions and implications for empirical, methodology and practice in Section 8.3. Section 8.4 acknowledges the research limitations. Section 8.5 offers directions for future research. Finally, the summary of this chapter is presented in Section 8.6.

8.2 Research Overview and Key Findings

The intention of this study was to provide a complete understanding of the explanatory power of Stock Micro-blogging sentiments in forecasting stock price behaviour and to determine whether they have the power to affect or alert investors' decision-making about specific types of traded stocks in the financial market. It highlights the role played by investor sentiments in determining assets pricing in the capital market. The following paragraphs summarise the general points that have evolved in the preceding seven chapters.

Chapter one is the introductory chapter to the entire research project. It provides an overview of and background to the study and defines the research problem. The chapter also discussed the motivations for conducting this research and

Chapter Eight: Conclusion

highlighted its relevance and significance. The Stock Micro-blogging service is one of the most popular investing community platforms where millions of investors share their investing opinions and thoughts on various traded securities of interest to them. As this is a relatively new source of financial information for opinion mining and sentiment classifications, a very limited number of studies have addressed this topic. One of the most important issues that render this research study significant is the need for automated textual analysis to extract online investor sentiments to examine whether or not these extracted sentiments have the power to predict stock market behavioural movements in the capital market. Previous empirical studies were conducted to validate the predictive ability of stock micro-blogging sentiments to forecast stock market prices; however, these studies mainly focused on the lead-lag relationships. Studying a simple lead-lag relationship was insufficient to provide a comprehensive and in-depth understanding of the role played by noise traders in the capital market, and such studies said little about the predictive power of sentiments (those extracted from the Stock Micro-blogging forum) to predict stock market behavioural movements in the financial market. This chapter also stated the aim and objectives of the research and provided an outline of the structure of this thesis.

Chapter two presents a critical review of the literature on the two fields of study covered in this thesis: finance and data mining. This chapter discusses the most relevant theories that serve as a framework to provide theoretical reasons for the assertions made in this thesis on the existence of the predictive ability of Stock Micro-blogging sentiments in predicting stock market behaviour; these are utilised as the foundation for this research. Behavioural finance research has provided evidence that the trading activities of noise traders and arbitrageurs will inevitably drive asset prices in the capital market. Three main online investment forums have also been discussed by identifying the related literature that highlights the underlying effects of each of these forums in forecasting different financial market indicators. Despite the numerous studies that provide some convincing evidence of the impact of such forums on the prediction of certain financial variables, there may be contradictory findings with regard to the predictive power of the stock messages obtained from these forums in predicting stock price behavioural movements. In order to extract the online sentiments from online text, various data mining and machine learning techniques were employed in this research. Considering the different assumptions and

Chapter Eight: Conclusion

biases of each of the machine learning techniques studied, one might expect varied results in the classification accuracy for sentiment detections in these online texts. This chapter also placed great emphasis on the role of each classifier when feature selection is employed in extracting the most relevant features from these online texts. It is argued that feature selection methods (filter and wrapper) improve the classification accuracy of classifiers. The two methods differ in terms of how the relevancy of the features is evaluated. Text mining plays a major role in online stock forums and the financial market. It was revealed that there is a lack of rigorous methodology in the analysis of financial indicators and a failure to implement multiple classifiers and feature selection tasks of text mining in predicting stock market behaviour in online stock forums.

Chapter three established the conceptual framework of this study, which designs the key variables, constructs the facts being studied and presumes the correlations between these studied variables. The conceptual framework focuses on the relationship between stock micro-blogging features and different financial indicator variables. The aim is to establish a framework for determining the predictive power of stock micro-blogging sentiments in predicting the behavioural movement of stock prices in the capital market. Different hypotheses were formulated to address the effect of stock micro-blogging features on various indicators of the financial market. This chapter provided an opportunity to gain a better understanding of the predictive ability of stock micro-blogging features in forecasting price movements in the stock market.

Chapter four presented the methodological approach applied in this study and justified its suitability for this particular research. This study employed a quantitative approach in the positivist paradigm and used secondary data as its main data resource. The main sources of the data were the Stock-related Micro-blogging website known as “StockTwits” and financial market data retrieved from Bloomberg. Data mining techniques and financial econometrics modelling are the two innovative methodologies employed in this research, and these are derived from the Information Systems and Empirical Finance disciplines, respectively. All studied variables are discussed in more detail in this chapter and the way in which each of these variables is measured and constructed is also explained. Following the measurement of the

Chapter Eight: Conclusion

variables, different statistical analysis techniques and empirical finance modelling used for empirical investigation of the research hypothesis were provided.

Chapter Five presented the initial findings of this research study in which different machine learning algorithms are used to perform the text-mining tasks for sentiment prediction of StockTwits data. The aim of this chapter was to provide an answer to one of the research questions: can text-mining techniques accurately predict sentiment analysis on StockTwits? To address this research question, a comparative analysis of sentiment automated classification of three different classifier algorithms, namely Naïve Bayes (NB), Random Forest (RandF) and Sequential Minimum Optimal (SMO), is conducted and evaluated. Since each of these three classifiers has its own bias, the sentiment classification accuracy may differ substantially. Therefore, different performance evaluation techniques are employed to determine the best text-mining techniques for sentiment prediction. Additionally, this chapter makes use of the feature selection method, which is considered one of the most essential tools of text mining as it focuses on the relevant features and omits irrelevant features from the dataset being processed. Two approaches to feature selection, i.e. filter and wrapper approaches are performed to investigate their effectiveness in improving the sentiment analysis tasks of the compared classifiers. The results of this chapter revealed that the RandF classifier algorithm outperforms the NB and SMO classifiers as it provides the highest sentiments accuracy and performs better with both the filter and wrapper approaches to feature selection. Therefore, the RandF classifier is the superior algorithm selected to perform the sentiment analysis tasks of StockTwits data. To provide a greater insight into the sentiment prediction tasks, this chapter presents novel approaches by combining text analysis tasks, feature selection and different machine learning algorithms (e.g. BN classifiers and wrapper approach in application 1, and Decision Tree C4.5 with filter approach in application 2). These two applications provide potential practical implications for investors and other market practitioners by offering them a real-time investing idea while providing them with an investing decision support mechanism that may help them to make better informed investments that maximise their investment returns in the capital market.

Chapter Six provides a preliminary analysis of StockTwits features and financial market variables including the basic descriptive statistics (i.e. mean, standard deviations, minimum and maximum) and some statistical analyses of the

Chapter Eight: Conclusion

distributions of StockTwits postings by ticker symbols, time, and day of the week. An exciting finding about the distributions of tweet postings by companies is that the top 10 stock tickers of the entire DJIA account for approximately 67% of all postings, thus indicating that these stocks are the most heavily discussed by investors and their peers in the StockTwits forum. It is also observed that the message posts show high activity during working days while there is a low volume of posts during weekends and that message posting is concentrated between 10:00am and 5:00pm; this suggests a high level of activity by day traders during market hours. This chapter also investigated the contemporaneous relationship between the tweet features and stock market variables by studying the pairwise correlations matrix and the contemporaneous regressions that address the independence of these relationships. The pairwise correlations analysis shows that the correlations between message volume and trading volume; bullishness and trading volume appear to be the most significant and robust and the magnitude of these relationships suggests that available information may exist in these relationships. The results of the contemporaneous regressions revealed that none of the StockTwits measures offers any statistically significant ability to predict stock market returns. While the trading volume and volatility regressions show more robust results with respect to the predictive ability of all StockTwits variables, the bullishness index shows the strongest effect in anticipating market variables (trading volume and volatility). Additionally, this chapter estimated the VAR model for each stock market variable independently. The VAR-return model showed the inability of all StockTwits features to predict the subsequent moves in returns. However, the Granger causality test showed that a significant causality is still running in the reverse direction for some tweet features such as message volumes and agreement. Both VAR models of trading volume and volatility imply that tweet features may contain predictive information in explaining volume and volatility. While the effect of StockTwits features (e.g. bullishness and agreement) has a more significant impact than in the reverse direction, the Granger causality test measured by Chi-square χ^2 implies that most of the relationships addressed in the VAR-volatility model show a much stronger effect of volatility on tweets features.

Chapter seven presents the findings of the roles played by noise traders in influencing stock market behaviour. In particular, the chapter explores the

Chapter Eight: Conclusion

asymmetrical behaviour of bullish and bearish sentiments on different stock market variables (return, volatility and trading volumes). The findings reveal that bullish (bearish) shifts in investor sentiment cause significant downward (upward) revision in volatility of return and are associated with higher (lower) stock returns. Regarding the trading volume, the results suggest a significant positive impact of investor sentiment on stock trading volume; i.e. the volume of trade increases (decreases) when investors become more bullish (bearish). This chapter also examines the asymmetrical response of investor sentiments in different states of the economy (i.e. bull/bear markets) where the findings reveal that positive returns trigger an increase in investor bullishness (decreased bearishness) in the bull market, while negative returns trigger a reduction in investor bullishness (and/or increased bearishness) in the bear market.

Using the Quantile regression approach, an in-depth investigation of the sentiment- return relationship is also conducted in this chapter and the results show that shift sentiment exerts opposite and heterogeneous effects on the two sides of the return distribution. The results confirm the existence of asymmetrical behaviour of bullish and bearish sentiments on stock return in which their impact is more pronounced at the extreme performances of stock return. Additionally, this chapter examined the non-linear relationship between investor disagreement and trading volume in two different market regimes. The empirical findings reveal that online divergence of the traders' opinion measure is proved to contain value-relevant information that is not yet reflected in securities' trading volumes and that relations are found to be symmetrical. A high level of disagreement among traders triggers an intense level of trading activities in both bull and bear markets, and the effect of disagreement on trading volume is more pronounced in the downwards market. While the results on volume portfolio strategies based on disagreement provide support for Miller's model (1977), in which high disagreement among traders earns lower expected returns, the impact of disagreement shows a distinct asymmetric effect on returns in the large and small traded stocks whereby lower (higher) returns are likely in the high disagreement portfolios for the large (small) traded stocks, respectively.

8.3 Research Contributions and Implications

This research study should prove to be a significantly rigorous and theoretically interesting field of research for both academics and practitioners through

Chapter Eight: Conclusion

its exploration of the impact of stock micro-blogging sentiments on stock market behaviour. This research has successfully integrated two areas of study, namely the data mining and financial econometrics fields, which are rapidly becoming ubiquitous in this area of academic research. The next subsections present and address the empirical, methodological and practical contributions of this study.

8.3.1 Empirical Contributions

This study's contribution to research is a decision support artefact using emerging social networks. The models and approach constructed herein may form the groundwork for future research where researchers and practitioners alike may find it fruitful to pay attention to the boom in financial blogs in order to understand the significant role of sentiments, especially micro-blogging sentiments, in predicting stock price behavioural movement in stock markets. This study contributes to two different research community groups in particular: the financial research community and the data mining community.

The theoretical investigation presented in this research contributes to the finance literature by strengthening ties with reference disciplines in tackling and addressing the on-going debate on the efficient market hypothesis (EMH) (Fama, 1970), random walk theory and behavioural finance theory. Overlooking behavioural factors in the traditional finance theory will fail to provide a complete picture of stock price prediction behaviour in the financial market. Behavioural finance economists have long been challenging the underlying assumptions of EMH. They argue that the market is not informationally efficient and little predictability may be possible. They also claim that not all investors are rational when making their investment choices and that the interactions between two types of heterogeneous agents, namely noise traders and arbitrageurs, will have a significant impact on stock price changes in the capital market. Empirical studies claim that the trading behaviour of noise traders in the capital market will create risk (called noise trader risk), which limits arbitrageurs' trading activities and deters them from betting against noise traders. The noise trader risk will drive the price level away from its fundamental value. Consequently, the noise trader has proved to have a significant effect on stock price movements; therefore, it is inherently important to account for the behavioural and psychological components in any assets pricing model that attempts to describe stock price

Chapter Eight: Conclusion

behaviour. An important implication of this research is that it provides significant evidences to investors, financial professionals and other market regulators that the StockTwits is significant and powerful investing forum and that online talk in this forum affects stock prices and confirms that there are inefficiencies in the stock market. This research thesis provides support for the significant yet arguable behavioural theories that highlight the role of irrational (noise) sentiments of investors in affecting asset returns. Hence, the proposed asset-pricing models may need to incorporate the role of such sentiments in financial markets to better explain asset returns.

Motivated by the noise trader risk model (DSSW, 1990), this research contributes to the literature by confirming the impact of online noise trader sentiment on stock price behaviour, as described in DSSW (1990). The DSSW model is considered a very relevant model in assets pricing as it predicts both the direction and magnitude of changes in noise trader sentiment. Conducting the empirical test of DSSW, researchers such as Brown (1999) and Graham and Harvey (1996) focused primarily on the impact of sentiment either on the mean or variance in assets returns, resulting in misspecifications and an incomplete picture. Following Lee et al. (2002), this research study addresses the impact of noise trader risk on both returns and the formation of volatility as suggested in DSSW (1990). More specifically, it investigates the four effects of noise trader risk: The ‘price-pressure’ and the ‘hold-more’ effects influence asset returns directly and reflect the contemporaneous shift in investor sentiments, while the ‘Friedman’ and ‘create-space’ effects reflect the long-run impact of noise traders indirectly on returns through the magnitude of sentiment changes on the future return volatility. This research is different than those of Lee et al. (2002) in three several ways. First, this study conducted on 30 stocks of DJIA index where the data are collected based on individual company level (data collected separately for each of these individual companies) rather than on market level (for DJIA index as whole). Second, the data used in this study is structured as panel data with company fixed effect (controlling for company specific characteristics such as size, etc.) unlike Lee et al. (2002) who employ time series data for individual indices (i.e., DJIA, S&P500 and NSDAQ). Third, contrasting Lee et al. (2002), who use Investor Intelligence survey data to measure investor sentiment, this study employs relatively new data for investor sentiment extracted from an online stock micro-

Chapter Eight: Conclusion

blogging forum called “StockTwits”. Although StockTwits data have only been recently used in different field of studies, such as mining data and sentiment lexicon (Oliveira et al., 2014) and stock market prediction (Oh and Sheng, 2011; Oliveira et al., 2013; Sprenger et al., 2014; Wang et al., 2014; Giannini et al., 2014;; Ranco et al., 2015), this study uses stock micro-blogging sentiment as a proxy for investor sentiment to investigate the impact of online noise trader risk on stock return, in light of the DSSW (1990) model of noise trader. This study is relatively new to the literature since previous studies of online investor sentiment have only focused on understanding whether online sentiments are leading or lagging the movements in the stock market (Tumarkin and Whitelaw, 2001; Antweiler and Frank, 2004b; Das and Chen, 2007; Oh and Sheng, 2011; Mao et al., 2012; Sprenger et al., 2014) whereas parts of this study went a step further by applying the “event study” methodology (Sprenger et al., 2014; Ranco et al., 2015). The aim is to confirm the existence of the noise trader’s role in online stock forums. By showing that online noise traders have the predictive ability to determine assets pricing, this research thesis also sheds light on new methodological data on how sentiments extracted from online forums might be used to analyse stock price behavioural movement in the financial market.

Additionally, this study contributes to the current literature by recognising the role of disagreement in trading volume in two different regimes/patterns of return, namely the bull and bear markets. The empirical evidence in this thesis establishes a link between two research areas: trading volume and investors’ disagreement. On the one hand, this thesis contributes to the research on trading volume by identifying one of the sources of the variation in stock market trading volumes, which is the investor divergence of opinion. On the other hand, it adds to the literature on investors’ behaviour by revealing how investor disagreement, by varying investors’ opinions, affects the aggregate volume of trade variation. The current literature in these two areas has reported that both trading volume and investor disagreement influence stock market returns. However, the connection between investor disagreement and trading volume proposed in this study allows an inclusive understanding in relation to investors’ behaviour (disagreement), trading volume and returns. Therefore, the empirical results in this thesis are able to successfully link the three respective variables of concern, namely the trading volume, the level of disagreement, and stock returns. Thus, this thesis demonstrates that disagreement is not only an important

Chapter Eight: Conclusion

factor in determining trading volumes, but it is also considered a very significant factor in influencing asset prices and returns in the capital market.

With regard to the Data Mining community, this research study takes a different approach by integrating text-mining techniques and feature selection methods with different machine learning algorithms such as Application A: Bayesian Network model performing the wrapper approach, and Application B: Decision Tree model performing the filter method, which are explained in elaborate detail in section 5.10. In Application (1), the aim was to investigate the interactions between the selected features and their ability to predict investors' sentiments quarterly over different periods of the year. In Application (2), the aim was to predict an intelligent trading support mechanism to screen out the most significant and profitable trading terms or combination of terms from StockTwits data that may help investors to make correct and accurate (selling, buying or holding) decisions in capital markets. In so doing, this research proposes a novel method that helps in the selection of an accurate set of relevant features, thus providing an insight into the relevancies present within the financial information used.

8.3.2 Methodological Contributions

This study is one of only a very few to have employed rigorous econometric modelling to address the issue of the asymmetric impact of investor sentiment by distinguishing between bullish and bearish sentiments on stock returns. As explained in Section 7.4.1, this study specifically compares the difference between the results of a traditional Ordinary Least Square (OLS) model and a quantile regression model of the relationship between bullishness (as a proxy for investor sentiments) and stock return. This quantile technique enables us to examine whether the relationship between sentiments and returns differs throughout the distribution of the dependent variable (i.e., stock returns). Unlike previous studies of sentiment-returns relations based on linear models that yield implications for the conditional mean of the dependent variable return given the explanatory variable sentiment, this study focuses on quantile regression (QR), which accounts for the behaviour of sentiments at various quantiles of returns. The quantile regression model provides more precise information about the distribution dispersion of the dependent variable while offering a more efficient estimate than the estimate obtained using OLS. Moreover, it provides

Chapter Eight: Conclusion

more robust results for coefficient estimates because QR is unresponsive to the effect of the outlier samples on the dependent variable and the fact that the distribution of the error term is not normally distributed.

Another important methodological contribution of this study, as noted in section 7.4.2, is that to date there has been no empirical research on how the divergence of online opinions (disagreement) affects asset prices and volume of trade in the capital market, apart from those carried out by Antweiler and Frank (2004b) and Sprenger et al. (2014). The above-mentioned studies used an online disagreement measure as a proxy for divergence of opinion, but they went no further than the simple lead-lag relationships with trading volumes. Therefore, this research has been undertaken to provide a rigorous methodology and detailed analysis of the impact of online divergence of opinion on trading volume. Additionally, this study considered both the linear and the non-linear effect of disagreement on trading volumes by empirically investigating the asymmetric effect of investors' opinion divergence on trading volume in two different regimes/patterns of return, namely bull and bear markets. The aim was to capture the asymmetry in the predictive power of investors' disagreement in trading volume in these two regimes separately. To the best of the author's knowledge, this is the first work to show the asymmetric impact of the differences in opinion on trading volumes in different states of the market: bull and bear markets. This study not only investigates the volume-disagreement relations but also provides an in-depth analysis of how these prospective relationships may be affected according to the type of news arriving in the market (good and bad news indicated by the bull and bear markets respectively).

8.3.3 Practical Contributions

This thesis contributes and adds value to two groups of practitioners who have been addressed in this study: investors (individual and institutional) and companies. First, for investors who are always looking for accurate methods to predict stock prices, this study makes a primary contribution providing real-time investing ideas by utilising stock micro-blogging sentiments. This offers the potential for practical applications, providing investors and their peers with an investment decision support mechanism. Second, this research presents a nascent approach by providing a robust methodology that could provide guidance to investors and other financial

Chapter Eight: Conclusion

professionals in constructing and rebalancing their investment portfolios (as noted in Section 5.10). This potentially offers guidelines to help investors and traders determine the correct time to invest in the market, what type of stocks or sectors to invest in, and which ones yield maximum returns on their investments.

Third, the predictive model of sentiment-returns using the quantile regression approach (as presented in chapter 7) has significant implications for investors' trading behaviour in financial markets, which might be explained by the interaction of two heterogeneous agents, namely noise traders and arbitrageurs, and through the concept of misperception. For example, noise traders' activities driven by their psychology, emotions, preferences and mistaken beliefs can affect the decisions of other investors, hence shifting the asset's value (up or down) from its mean, which may create risk and limit arbitrageurs' ability to beat them. While the arbitrageurs' actions are limited in the short run, this limitation will diminish in the long run when prices' deviation from mean levels becomes sufficiently extreme (high and low). Hence, the finding that sentiment plays a notable role in predicting stock returns in extreme market conditions may be due to noise traders' overreaction to good and bad news, i.e. market growth and recession, which can cause price levels and risk to deviate wildly from expected levels that would have actually been set by the news. In spite of the prices' deviations from mean values and the limits on arbitrage, this predictive modelling provides an investment decision support mechanism that helps prevent arbitrageurs from taking immediate action to avoid falling into a possible trap of misperception. The misperception about the assets' risk causes noise traders to follow each other in terms of selling (buying) risky assets just when noise traders are doing so. Additionally, our predictive modelling provides arbitrageurs with appropriate market timing regarding their contrarian trading strategies as the misperception of noise traders becomes even more extreme when new fundamental information arrives unexpectedly after an arbitrageur has taken her/his initial position. Overall, the results of this thesis provide regulators with confirmation that the StockTwits forum affects stock prices. These results also confirm to investors that inefficiencies exist in the stock market. This suggests that corporate managers, especially those in small firms, should monitor the StockTwits forum.

On the other hand, the adoption of new media would achieve a competitive advantage for a company and help it to keep its stakeholders informed and

Chapter Eight: Conclusion

empowered. Furthermore, for quality and profitability purposes, companies might also monitor customer posts regarding their products and services to maintain the high quality of those products or services. Moreover, companies and managers might choose to disseminate their financial reporting information and/or advertise with postings deemed to have higher predictive value. Text Mining techniques are also aimed at finding Business Intelligence solutions to help companies remain competitive in the market (Bolasco et al., 2005).

8.4 Research Limitations

Although this research provides novel and significant insights and draws valuable lessons with regard to the explanatory power of Stock Micro-blogging sentiments in predicting stock market behaviour and to the role played by investor sentiments in pricing assets in the capital market, it is not, like others, without certain limitations. Therefore, this section brings up some of these limitations, thereby opening up fruitful avenues for future research. The limitations of this study are divided into two types - method and data limitations - as follows:

8.4.1 Method limitations

As with any other fields of study, the research design employed in this study is not come out without certain methodological limitations. A number of limitations to the research methodology should be noted.

- Inter-coding agreement methods employ in this study may require a considerable amount of time and effort to train a second coder to manually classify StockTwits posts in terms of reaching exact agreement on the coding scheme.
- In data mining techniques, extracting the most relevant features by adopting the wrapper approach to feature selection is a time-consuming task that will require more time and a faster machine to achieve.
- Using the Harvard IV dictionary from General Inquirer may limit the words used in the manual classifications, which may then affect the classification accuracy rate in sentiment detection. Therefore, it may be necessary to extend

Chapter Eight: Conclusion

the dictionary list for sentiment classifications under each sentiment class (sell, buy and hold) to increase the level of sentiment accuracy. An expanded lexicon from the training data might help increase the classification accuracy rate in sentiment detection.

8.4.2 Data limitations

There are some limitations to the data used in this research study, that are worth to be noted and addressed, which then might open up fruitful avenues for future research.

- Non-English language tweets on the StockTwits platform were ignored; however, they might have had an impact or effect on stock price movements in the capital market.
- There are missing observations (silent periods where no tweets are posted on those dates). Following Antweiler and Frank (2004b) on the Internet message board, all silent periods are given a value of zero. Although this method is the most widely used approach to dealing with silent periods in all studies, the inclusion of the zeros may bias some of the tweet measures that are considered the most critical measures obtained from the tweet postings (e.g. bullishness index and the level of disagreement).
- The sample period is short (one year's data is a very small amount due to the difficulty of dealing with millions of tweet messages on all the companies registered on StockTwits all over the globe). One of the disadvantages associated with the shorter study period is that the short time period prevents the researcher from using some econometric modelling whose usage is highly sensitive to the length of time covered. For example, in this research study, the intention was to employ a member of the family of GARCH Models to estimate the stock return volatility; however, the researcher was unable to find an ARCH effect considering this short study period where, in the normal setting, the longer period covered the higher probability of passing the ARCH effect test.
- Considering the large dataset collected from the StockTwits website, a very challenging task faced by the researcher was to distinguish the source of the tweet messages; i.e. were they contributed by private investors or institutional

Chapter Eight: Conclusion

investors? Given that individual and intuitional investors differ in their needs and practices, one might expect the messages tweeted by private investors to have a different effect on the stock market from those tweeted by institutional investors.

8.5 Directions for Future Research

While this research study provides empirical investigations of the predicted ability of online stock forums in predicting stock market behaviour in financial market, several beneficial areas of future research, however, remain to be explored. There are many avenues for future research which are addressed as following:

- This research studies the predictive relationship of stock micro-blogging sentiments in predicting stock market behaviour using a daily granularity of analysis. However, in studying the timeline of StockTwits, an analysis can be performed to determine whether the time of day at which a tweet is sent has a sentiment effect and whether there are any days of the week or month that have a more positive or negative correlation with the movement in the stock market. Moreover, the analysis of sentiment might be extended to determine and report sentiment in almost real time (intraday data: e.g. hourly data) to allow investors and traders to decide in which stocks or sectors they should invest throughout the trading day.
- The models and approaches used in this research study may provide insights for future research studies seeking to understand the predictive value of stock micro-blogging sentiments. Although the model adopted in this thesis incorporates these sentiments as a parameter bound to give superior prediction at the micro-economic level, it might be accepted that these sentiments will also drive macro-economic movements in the market. With the help of data mining and various textual analysis tools, researchers are able to extract sentiments from the opinions of large numbers of public users. A considerable number of studies (such as Bollen et al., 2010; Zhang, 2009; Antweiler and Frank, 2004b; Das and Chen, 2007; Sprenger et al., 2014) have addressed these issues and produced interesting results. “The wisdom of the crowd”, which is defined as the process of considering the collective opinion of a group of individuals, furnished with data mining and machine learning

Chapter Eight: Conclusion

applications, can automatically provide estimations and predictions on a variety of subjects, such as stock market predictions (Sprenger et al., 2014), predicting box office revenues (Asur and Huberman, 2010), swine flu spread prediction (Ritterman et al., 2009), and predicting disaster news spread (Doan et al., 2012).

- An examination of the non-linear model in studying the reactions of investor sentiment to stock returns in two different states of the economy (bull and bear markets) may provide direction for future studies that might investigate the investor sentiment reactions to changes in volatility considering the high- and low-volatility periods. Since this study shows a significant relationship between volatility and the bullishness measure (used as a proxy for investor sentiment), this suggests that variability in stock prices may cause a great deal of variation in investor sentiment. The study also proved the predictive ability of stock volatility, which contains valuable information for explaining current investor bullishness. The significant relationship indicates the existence of an asymmetric effect of market volatility on the bullishness of investors. One way of recognising the high- and low-volatility regimes is through the estimations of the regime probabilities of the index return data to identify two regime classifications based on the smoothed probability of high- and low-volatility periods. As a result, two indicator variables are created for the high- and low-volatility periods while two interaction terms for each period are inserted into the sentiment equation, as in the case of the bull and bear markets. Hence, to check the asymmetric volatility effect on investor bullishness, the coefficients of the interaction terms during the high- and low-volatility regimes in the sentiment equation should be evaluated.
- Another extension of this study would be to consider training the model on StockTwits of each company ticker making up the Dow index, separately over a longer period of time to explore and investigate the firm-specific terms and how different terms and/or combinations of terms may interact and interrelate in each ticker rather than considering the market index as a whole. This is expected to help improve the performance of ticker sentiment prediction.

Chapter Eight: Conclusion

- Comparative study cross-indices might be conducted to compare the predictive power of stock micro-blogging sentiments and to determine whether there are substantial changes in the degree of predictive ability of cross-indices while investigating the factors resulting in such changes, if any. In addition, future research into how Twitter sentiments might be used to predict movements of a particular stock or sector may yield promising insights into potential practical applications. This could, potentially, provide a decision support mechanism for investors and traders to use while attempting to determine whether to invest in a particular stock or sector.
- Extended datasets should include more data covering longer periods of time. Considering larger periods of time is believed likely to produce more significant results.
- Much work remains to be done in studying the effects of sentiments on the stock market. Research needs to be directed to cover non-US markets. A considerable number of research studies in this area have been conducted using US market data, such as those by Antweiler and Frank (2004b) and Sprenger et al. (2014). Considering other stock markets apart from US market indices may provide greater insights into the impact of investor sentiment on stock markets and may reveal whether these impacted powers vary across indices of different countries.

8.6 Chapter Summary

This final chapter provided a conclusive summary of the results and discussions of the research presented in this thesis. First, an overall summary of the research findings has been presented. Second, the theoretical, practical and methodological contributions and implications of this research have been highlighted. Third, the research limitations in terms of methods and data have been acknowledged. Finally, fruitful directions for future studies have been identified.

Having arrived at the end of this thesis, it is important to note that this study was conducted in a relatively new research domain covering two fields of study: Data Mining and Empirical Finance. This study presents a novel approach by combining text-mining techniques with empirical finance models to explore the predictive relationships of the StockTwits sentiments on the stock market while investigating the

Chapter Eight: Conclusion

role played by investor sentiments and other tweet features in determining asset prices and their movements in the capital market. Thus, the researcher has taken the first step towards providing a more comprehensive understanding and analysis of the significant role played by stock micro-blogging sentiments in predicting stock price behavioural movements. This may yield promising insights into the potential provision of an investment support mechanism for analysts, investors and their peers. The results and discussions of this research may arguably contribute to theoretical understanding and should be of direct practical value to researchers and practitioners alike. The researcher hopes that this study has added significant value to the field and encourages future researchers to undertake fruitful work exploring in greater depth this burgeoning and interesting area of research.

References

REFERENCES

Anonymous 2011. Research and Markets: The StockTwits Edge: 40 Actionable Trade Set-Ups from Real Market Pros. *Business Wire*.

Abbasi, A., Chen, H. and Salem, A. 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. *ACM Transactions on Information Systems (TOIS)*, **26**(3), pp. 12.

Abraham, R., Simha, J.B. and Iyengar, S. 2007. Medical data mining with a new algorithm for Feature Selection and Naïve Bayesian classifier, *Information Technology, (ICIT 2007)*. *10th International Conference on 2007*, IEEE, pp. 44-49.

Acemoglu, D., Ozdaglar, A. and Parandehgheibi, A. 2010. Spread of (mis) information in social networks. *Games and Economic Behavior*, **70**(2), pp. 194-227.

Aitken, B. 1998. Have Institutional Investors Destabilized Emerging Markets? *Contemporary Economic Policy*, **16**(2), pp. 173-184.

Aizerman, M., Braverman, E. and Rozonoer, L. 1965. The probability problem of pattern recognition learning and the method of potential functions(Probability of pattern recognition learning by robots, using algorithm based on potential function method). *Automation and Remote Control*, **25**, pp. 1175-1190.

Akaike, H. 1974. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, **19**(6), pp. 716-723.

Alagidede, P. and Panagiotidis, T. 2012. Stock returns and inflation: Evidence from quantile regressions. *Economics Letters*, **117**(1), pp. 283-286.

Alizadeh, S., Brandt, M.W. and Diebold, F.X. 2002. Range-based estimation of stochastic volatility models. *The Journal of Finance*, **57**(3), pp. 1047-1091.

Amari, S. and Wu, S. 1999. Improving support vector machine classifiers by modifying kernel functions. *Neural Networks*, **12**(6), pp. 783-789.

Andersen, T.G., Bollerslev, T., Diebold, F.X. and Ebens, H. 2001. The distribution of realized stock return volatility. *Journal of Financial Economics*, **61**(1), pp. 43-76.

Andersen, T.G., Bollerslev, T., Diebold, F.X. and Labys, P. 2003. Modeling and forecasting realized volatility. *Econometrica*, **71**(2), pp. 579-625.

Andrade, A.D. 2009. Interpretive research aiming at theory building: Adopting and adapting the case study design. *The Qualitative Report*, **14**(1), pp. 42-60.

References

- Antoniou, C., Doukas, J.A. and Subrahmanyam, A. 2013. Cognitive dissonance, sentiment, and momentum. *Journal of Financial and Quantitative Analysis*, 48(01), pp.245-275.
- Antweiler, W. and Frank, M. 2002. Internet stock message boards and stock returns. *University of British Columbia Working Paper*, pp. 1-7.
- Antweiler, W. and Frank, M.Z. 2004a. Does talk matter? Evidence from a broad cross section of stocks. *University of British Columbia Working Paper*.
- Antweiler, W. and Frank, M.Z. 2004b. Is all that talk just noise? The information content of Internet stock message boards. *The Journal of Finance*, 59(3), pp. 1259-1294.
- Artstein, R. and Poesio, M. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4), pp. 555-596.
- Asur, S. and Huberman, B.A. 2010. Predicting the future with social media, *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on 2010*, IEEE, pp. 492-499.
- Bacidore, J., and M.L. Lipson, 2001, “The Effects of Opening and Closing Procedures on the NYSE and NASDAQ”, Working Paper, NYSE and University of Georgia.
- Baek, I., Bandopadhyaya, A. and DU, C. 2005. Determinants of market-assessed sovereign risk: Economic fundamentals or market risk appetite? *Journal of International Money and Finance*, 24(4), pp. 533-548.
- Baik, B., Cao, Q., Choi, S. and Kim, J.M., 2015. Local Twitter Activity and Stock Returns. Working Paper
- Bailey, C.A., 2007. *A guide to qualitative field research*. London :Sage Publications.
- Baillie, R.T., Bollerslev, T. and Mikkelsen, H.O., 1996. Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 74(1), pp. 3-30.
- Baker, M. and Stein, J.C. 2004. Market liquidity as a sentiment indicator. *Journal of Financial Markets*, 7(3), pp. 271-299.
- Baker, M. and Wurgler, J. 2006. Investor Sentiment and the Cross-Section of Stock Returns. *The Journal of Finance*, 61(4), pp. 1645-1680.
- Baker, M. and Wurgler, J. (2007). Investor sentiment in the stock market. *Journal of Economic Perspectives*, 21 (2): 129–151.
- Baker, M., Wurgler, J. and Yuan, Y. 2012. Global, local, and contagious investor sentiment. *Journal of Financial Economics*, 104(2), pp. 272-287.

References

- Ballou, D.P. and Pazer, H.L. 1995. Designing information systems to optimize the accuracy-timeliness tradeoff. *Information Systems Research*, **6**(1), pp. 51-72.
- Balvers, R., Wu, Y. and Gilliland, E. 2000. Mean reversion across national stock markets and parametric contrarian investment strategies. *The Journal of Finance*, **55**(2), pp. 745-772.
- Bamber, L.S., Barron, O.E. and Stober, T.L. 1999. Differential interpretations and trading volume. *Journal of financial and Quantitative Analysis*, **34**(03), pp. 369-386.
- Banerjee, S. and Kremer, I. 2010. Disagreement and learning: Dynamic patterns of trade. *The Journal of Finance*, **65**(4), pp. 1269-1302.
- Banz, R.W. 1981. The relationship between return and market value of common stocks. *Journal of Financial Economics*, **9**(1), pp. 3-18.
- Barakat, N.H. and Bradley, A.P. 2007. Rule extraction from support vector machines: a sequential covering approach. *Knowledge and Data Engineering, IEEE Transactions on*, **19**(6), pp. 729-741.
- Barber, B.M. and Odean, T. 2008. All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *Review of Financial Studies*, **21**(2), pp. 785-818.
- Barber, B.M. and Odean, T. 2001. The Internet and the investor. *Journal of Economic Perspectives*, pp. 41-54.
- Barber, B.M. and Odean, T. 2000. Trading is hazardous to your wealth: The common stock investment performance of individual investors. *Journal of Finance*, , pp. 773-806.
- Barberis, N. and Thaler, R. 2003. A survey of behavioral finance. *Handbook of the Economics of Finance*, **1**, pp. 1053-1128.
- Basak, S. 2005. Asset pricing with heterogeneous beliefs. *Journal of Banking & Finance*, **29**(11), pp. 2849-2881.
- Basu, S. 1977. Investment performance of common stocks in relation to their price-earnings ratios: A test of the efficient market hypothesis. *The journal of Finance*, **32**(3), pp. 663-682.
- Baur, D.G., Dimpfl, T. and Jung, R.C. 2012. Stock return autocorrelations revisited: A quantile regression approach. *Journal of Empirical Finance*, **19**(2), pp. 254-265.
- Benartzi, S. and Thaler, R.H. 2001. Naive diversification strategies in defined contribution saving plans. *American economic review*, pp. 79-98.
- Bennett, K.P. and Mangasarian, O.L. 1992. Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software*, **1**(1), pp. 23-34.

References

- Benos, A.V. 1998. Aggressiveness and survival of overconfident traders. *Journal of Financial Markets*, **1**(3), pp. 353-383.
- Berg, B.L. and Lune, H. 2014. *Qualitative research methods for the social sciences*. 8th, Pearson new international edn. Harlow: Pearson.
- Bergmeir, C. and Benítez, J.M. 2012. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, **191**, pp. 192-213.
- Berkman, H., Dimitrov, V., Jain, P.C., Koch, P.D. and Tice, S. 2009. Sell on the news: Differences of opinion, short-sales constraints, and returns around earnings announcements. *Journal of Financial Economics*, **92**(3), pp. 376-399.
- Bernard, H.R. and Ryan, G. 1998. Text analysis. *Handbook of methods in cultural anthropology*, pp. 613.
- Bettman, J.L., Hallett, A.G. and Sault, S. 2011. Rumortrage: Can investors profit on takeover rumors on internet stock message boards? *Finance and Corporate Governance Conference 2011*.
- Black, F. 1986. Noise. *The Journal of Finance*, **41**(3), pp. 529-543.
- Blair, D.C. and Maron, M. 1990. Full-text information retrieval: Further analysis and clarification. *Information Processing and Management*, **26**(3), pp. 437-447.
- Blair, D.C. 1979. Information Retrieval, 2nd ed. C.J. Van Rijsbergen. London: Butterworths; *Journal of the American Society for Information Science*, **30**(6), pp. 374-375.
- Bloomberg L.P. 2013. *Stock price data for DJIA Index. 03/04/2012 to 05/04/2013*. Retrieved April 14, 2013 from Bloomberg terminal.
- Blum, A.L. and Langley, P. 1997. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, **97**(1), pp. 245-271.
- Bodurtha, J.N., Kim, D. and Lee, C.M. 1995. Closed-end country funds and US market sentiment. *Review of Financial Studies*, **8**(3), pp. 879-918.
- Boeije, H.R., 2009. *Analysis in qualitative research*. London: Sage Publication.
- Bohl, M.T. and Henke, H. 2003. Trading volume and stock market volatility: The Polish case. *International Review of Financial Analysis*, **12**(5), pp. 513-525.
- Bolasco, S., Canzonetti, A., Capo, F.M., Della Ratta-Rinaldi, F. and Singh, B.K. 2005. Understanding text mining: A pragmatic approach. *Knowledge Mining*. Springer, pp. 31-50.
- Bollen, J., Mao, H. and Zeng, X. 2011. Twitter mood predicts the stock market. *Journal of Computational Science*, **2**(1), pp. 1-8.

References

- Bollen, J., Pepe, A. and Mao, H. (2010) Modeling public mood and emotion: Twitter sentiment and socioeconomic phenomena, Proceedings from the 19th International World Wide Web Conference, Raleigh, North Carolina.
- Bondt, W.F. and Thaler, R. 1985. Does the stock market overreact? *The Journal of finance*, **40**(3), pp. 793-805.
- Bradley, A.P. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, **30**(7), pp. 1145-1159.
- Brooks, C., 2008. *Introductory econometrics for finance*. Cambridge university press.
- Brown, G.W. 1999. Volatility, sentiment, and noise traders. *Financial Analysts Journal*, **55**(2), pp. 82-90.
- Brown, G.W. and Cliff, M.T. 2005. Investor Sentiment and Asset Valuation. *The Journal of Business*, **78**(2), pp. 405-440.
- Brown, G.W. and Cliff, M.T. 2004. Investor sentiment and the near-term stock market. *Journal of Empirical Finance*, **11**(1), pp. 1-27.
- Brown, J.S. and Duguid, P. 2002. *The social life of information*. Harvard Business Press.
- Bryman, A. 2012. *Social research methods*. Oxford university press.
- Bryman, A. and Cramer, D., 2001. *Quantitative data analysis with SPSS release 10 for Windows: a guide for social scientists*. Routledge.
- Buchinsky, M. 1998. Recent advances in quantile regression models: a practical guideline for empirical research. *Journal of human resources*, pp. 88-126.
- Burrell, G. and Morgan, G. 1979. *Sociological paradigms and organisational analysis: elements of the sociology of corporate life*. London: Heinemann Educational.
- BusinessWeek, 2009, StockTwits may change how you trade, *BusinessWeek, Online Edition* (author Max Zeledon), February 11.
- Butler, K.C. and Malaikah, S., 1992. Efficiency and inefficiency in thinly traded stock markets: Kuwait and Saudi Arabia. *Journal of Banking & Finance*, **16**(1), pp. 197-210.
- Campbell, J.Y. and Kyle, A.S. 1993. Smart money, noise trading and stock price behaviour. *The Review of Economic Studies*, **60**(1), pp. 1-34.
- Campbell, J.Y. and Shiller, R.J. 1988. The dividend-price ratio and expectations of future dividends and discount factors. *Review of Financial Studies*, **1**(3), pp. 195-228.
- Campbell, J.Y. and Shiller, R.J. 1988. Stock prices, earnings, and expected dividends. *The Journal of Finance*, **43**(3), pp. 661-676.

References

- Campbell, J.Y. and Shiller, R.J. 1986. *Cointegration and tests of present value models. Journal of political Economy*, **95**, 1062-1088.
- Cavana, R., Delahaye, B.L. and Sekeran, U. 2001. *Applied business research: Qualitative and quantitative methods*. John Wiley and Sons Australia.
- Cawley, G.C. and Talbot, N.L. 2003. Efficient leave-one-out cross-validation of kernel Fisher discriminant classifiers. *Pattern Recognition*, **36**(11), pp. 2585-2592.
- Chan, L.K., Hamao, Y. and Lakonishok, J. 1991. Fundamentals and stock returns in Japan. *The Journal of Finance*, **46**(5), pp. 1739-1764.
- Chang, C. 2007. A study of applying data mining to early intervention for developmentally-delayed children. *Expert Systems with Applications*, **33**(2), pp. 407-412.
- Chang, P., Fan, C. and Lin, J. 2011. Trend discovery in financial time series data using a case based fuzzy decision tree. *Expert Systems with Applications*, **38**(5), pp. 6070-6080.
- Chang, T. 2011. A comparative study of artificial neural networks, and decision trees for digital game content stocks price prediction. *Expert Systems with Applications*, **38**(12), pp. 14846-14851.
- Chang, Y. Y., Faff, R. & Hwang, C. Y. 2011. Local and global sentiment effects, and the role of legal, trading and information environments. Working Paper.
- Charoenrook, A. 2005. Does sentiment matter? *Working Paper*, Vanderbilt University.
- Chen, H., De, P., Hu, Y. and Hwang, B. 2012. Customers as Advisors: The Role of Social Media in Financial. *Management Science*, **54**(3), pp. 477-491.
- Chen, J., Hong, H. and Stein, J.C. 2002. Breadth of ownership and stock returns. *Journal of Financial Economics*, **66**(2), pp. 171-205.
- Chen, P., Lin, C. and Scholkopf, B. 2005. A tutorial on v-support vector machines. *Applied Stochastic Models in Business and Industry*, **21**(2), pp. 111-136.
- Chen, R. and Hsieh, C. 2006. Web page classification based on a support vector machine using a weighted vote schema. *Expert Systems with Applications*, **31**(2), pp. 427-435.
- Chen, S., Lux, T. and Marchesi, M. 2001. Testing for non-linear structure in an artificial financial market. *Journal of Economic Behavior & Organization*, **46**(3), pp. 327-342.
- Chen, W. and Hirschheim, R. 2004. A paradigmatic and methodological examination of information systems research from 1991 to 2001. *Information systems journal*, **14**(3), pp. 197-235.
- Chevapatrakul, T. 2015. Monetary environments and stock returns: International evidence based on the quantile regression technique. *International Review of Financial Analysis*, **38**, pp. 83-108.

References

- Chien, C., Wang, W. and Cheng, J. 2007. Data mining for yield enhancement in semiconductor manufacturing and an empirical study. *Expert Systems with Applications*, **33**(1), pp. 192-198.
- Chiu, S., Chen, C. and Lin, T. 2008. Using support vector regression to model the correlation between the clinical metastases time and gene expression profile for breast cancer. *Artificial Intelligence in Medicine*, **44**(3), pp. 221-231.
- Cho, S., Asfour, S., Onar, A. and Kaundinya, N. 2005. Tool breakage detection using support vector machine learning in a milling process. *International Journal of Machine Tools and Manufacture*, **45**(3), pp. 241-249.
- Chordia, T., Roll, R. and Subrahmanyam, A. 2001. Market liquidity and trading activity. *The Journal of Finance*, **56**(2), pp. 501-530.
- Chuang, C., Kuan, C. and Lin, H. 2009. Causality in quantiles and dynamic stock return–volume relations. *Journal of Banking and Finance*, **33**(7), pp. 1351-1360.
- Chrysostomou, K.A. 2008. The Role of Classifiers in Feature Selection: Number vs Nature, PhD Thesis, Brunel University, UK.
- Chung, S., Hung, C. and Yeh, C. 2012. When does investor sentiment predict stock returns? *Journal of Empirical Finance*, **19**(2), pp. 217-240.
- Churchill, G. 1999. Marketing research: methodological foundations. 5th ed. Chicago: The Dryden Press.
- Claburn, T. 2009 “Twitter growth surges 131% in March” *Information Week*. Retrieved 25 Oct, 2013 from http://www.informationweek.com/news/internet/social_network/showArticle.jhtml?articleID=2165_00968
- Clark, P.K. 1973. A subordinated stochastic process model with finite variance for speculative prices. *Econometrica: Journal of the Econometric Society*, pp. 135-155.
- Clarkson, P.M., Joyce, D. and Tutticci, I. 2006. Market reaction to takeover rumour in Internet Discussion Sites. *Accounting and Finance*, **46**(1), pp. 31-52.
- Cohen, A. 1990. A Cross-Cultural Study of the Effects of Environmental Unpredictability on Aggression in Folktales. *American Anthropologist*, **92**(2), pp. 474-481.
- Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, **20**(1), pp. 37-46.
- Cohen, L., Manion, L. and Morrison, K. 2000. Research Methods in Education [5 th edn] London: Routledge Falmer. *Teaching in Higher Education*, **41**.
- Collis, J. and Hussey, R. 2013. *Business research: A practical guide for undergraduate and postgraduate students*. Palgrave Macmillan.

References

- Comiskey, E.E., Walkling, R.A. and Weeks, M.A. 1987. Dispersion of expectations and trading volume. *Journal of Business Finance and Accounting*, **14**(2), pp. 229-239.
- Cooper, D. and Schindler, P. 2001. Business research methods. *The Irwin/McGraw-Hill series, Operations and decision sciences*.
- Cootner, P.H. 1964. The random character of stock market prices. Cambridge, Mass M.I.T. Press.
- Copeland, T.E. 1976. A Model of Asset Trading under the Assumption of Sequential Information Arrival. *Journal of Finance*, **31**(4), pp. 1149-1168.
- Cortes, C. and Vapnik, V. 1995. Support vector machine. *Machine Learning*, **20**(3), pp. 273-297.
- Creswell, J.W. 2012. *Qualitative inquiry and research design: Choosing among five approaches*. London:Sage Publication.
- Cristianini, N. and Shawe-Taylor, J. 2000. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.
- Czekaj, T., Wu, W. and Walczak, B. 2008. Classification of genomic data: Some aspects of feature selection. *Talanta*, **76**(3), pp. 564-574.
- Danthine, J. and Moresi, S. 1993. Volatility information and noise trading. *European Economic Review*, **37**(5), pp. 961-982.
- Das, S.R. and Chen, M.Y. 2007. Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science*, **53**(9), pp. 1375-1388.
- Das, S.R. and Sisk, J. 2005. Financial Communities. *The Journal of Portfolio Management*, **31**(4), pp. 112-123.
- Das, S. and Chen, M. 2001. Yahoo! for Amazon: Extracting market sentiment from stock message boards, *Proceedings of the Asia Pacific finance association annual conference (APFA) 2001*, Bangkok, Thailand, pp. 43.
- Das, S., Martinez-Jerez, A. and Tufano, P. 2005. eInformation: A clinical study of investor discussion and sentiment. *Financial Management*, **34**(3), pp. 103-137.
- Das, S. J., and J. Sisk, 2005, "Financial Communities," *Journal of Portfolio Management*, **31**, 112–123.
- Dash, M., Choi, K., Scheuermann, P. and Liu, H. 2002. Feature selection for clustering-a filter solution, *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on 2002*, IEEE, pp. 115-122.
- Dash, M. and Liu, H. 1997. Feature selection for classification. *Intelligent data analysis*, **1**(3), pp. 131-156.

References

- Davis, A.K., Piger, J.M. and Sedor, L.M. 2012. Beyond the Numbers: Measuring the Information Content of Earnings Press Release Language. *Contemporary Accounting Research*, **29**(3), pp. 845-868.
- Davison, R. 1998. An action research perspective of group support systems: How to improve meetings in Hong Kong. *Unpublished PhD Dissertation, City University of Hong Kong*.
- De Bondt, W.P. 1993. Betting on trends: Intuitive forecasts of financial risk and return. *International Journal of Forecasting*, **9**(3), pp. 355-371.
- Bondt, W.F. and Thaler, R., 1985. Does the stock market overreact?. *The Journal of finance*, **40**(3), pp.793-805.
- De Long, J.B., Shleifer, A., Summers, L.H. and Waldmann, R.J. 1991. The survival of noise traders in financial markets. *Journal of Business*, pp. 1-19.
- De Long, J.B., Shleifer, A., Summers, L.H. and Waldmann, R.J. 1990. Noise Trader Risk in Financial Markets. *The Journal of Political Economy*, **98**(4), pp. 703-738.
- De Santis, Giorgio, Bruno Gerard and Pierre Hillion, 1999, The Single European Currency and World Equity Markets, in *European Markets with a Single Currency*, edited by Jean Dermine and Pierre Hillion, Oxford, University Press.
- De Souza, J.T., Matwin, S. and Japkowicz, N. 2006. Parallelizing feature selection. *Algorithmica*, **45**(3), pp. 433-456.
- Deb, S.G. 2012. Value versus growth: Evidence from India. *IUP Journal of Applied Finance*, **18**(2), pp. 48.
- Delort, J., Arunasalam, B., Leung, H. and Milosavljevic, M. 2012. The impact of manipulation in Internet stock message boards. *International Journal of Banking and Finance*, **8**(4), pp. 1.
- DeMarzo, P. 2003. M., Dimitri Vayanos and Jeffrey Zwiebel: Persuasion Bias. *Social Influence, and Unidimensional Opinions*, *Quarterly Journal of Economics*, **115**(3).
- Dewally, M. 2003. Internet investment advice: Investing with a rock of salt. *Financial Analysts Journal*, **59**(4), pp.65-77.
- Diamond, D.W. and Verrecchia, R.E. 1987. Constraints on short-selling and asset price adjustment to private information. *Journal of Financial Economics*, **18**(2), pp. 277-311.
- Dickey, D.A. and Fuller, W.A. 1979. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association*, **74**(366a), pp. 427-431.
- Diether, K.B., Malloy, C.J. and Scherbina, A. 2002. Differences of opinion and the cross section of stock returns. *The Journal of Finance*, **57**(5), pp. 2113-2141.

References

- Doan, S., Vo, B.H. and Collier, N. 2012. An analysis of Twitter messages in the 2011 Tohoku Earthquake. *Electronic Healthcare*. Springer, pp. 58-66.
- Dodds, P.S. and Danforth, C.M. 2010. Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies*, **11**(4), pp. 441-456.
- Duflo, E. and Saez, E. 2002. Participation and investment decisions in a retirement plan: The influence of colleagues' choices. *Journal of public Economics*, **85**(1), pp. 121-148.
- Easterby-Smith, M., Thorpe, R. and Jackson, P.R. 2012. *Management research*. London:Sage Publication.
- Eisenhardt, K.M. and Graebner, M.E. 2007. Theory building from cases: Opportunities and challenges. *Academy of management journal*, **50**(1), pp. 25-32.
- Eldabi, T., Irani, Z., Paul, R.J. and Love, P.E. 2002. Quantitative and qualitative decision-making methods in simulation modelling. *Management Decision*, **40**(1), pp. 64-73.
- Ellison, N.B. 2007. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, **13**(1), pp. 210-230.
- Engelberg, J. 2008. Costly information processing: Evidence from earnings announcements, Working Paper, University of North Carolina.
- Engle, R.F. and Patton, A.J. 2001. What good is a volatility model. *Quantitative finance*, **1**(2), pp. 237-245.
- Fama, E.F. 1991. Efficient capital markets: II. *The journal of finance*, **46**(5), pp. 1575-1617.
- Fama, E.F. 1970. Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, **25**(2), pp. 383-417.
- Fama, E.F., 1965a. Random walks in stock market prices. *Financial Analysts Journal*, **21**(5), pp. 55-59.
- Fama, E.F. 1965b. The behavior of stock-market prices. *The journal of Business*, **38**(1), pp. 34-105.
- Fama, E.F. and French, K.R. 1988. Permanent and temporary components of stock prices. *The Journal of Political Economy*, pp. 246-273.
- Fama, E.F. and French, K.R., 1989. Business conditions and expected returns on stocks and bonds. *Journal of Financial Economics*, **25**(1), pp. 23-49.
- Fan, W., Wallace, L., Rich, S. and Zhang, Z. 2006. Tapping the power of text mining. *Communications of the ACM*, **49**(9), pp. 76-82.

References

- Fawcett, T. 2006. An introduction to ROC analysis. *Pattern Recognition Letters*, **27**(8), pp. 861-874.
- Feinerer, I. 2008. An introduction to text mining in R. *R News*, **8**(2), pp. 19-22.
- Feinerer, I., Hornik, K. and Meyer, D. 2008. Text mining infrastructure in R. *Journal of Statistical Software*, **25**(5), pp. 1-54.
- Feng, Y., Chen, R. and Basset, G. 2008. Quantile momentum. *Statistics and its interface*, **1**, pp. 243-254.
- Fisher, K. L., Statman, M., 2000. Investor sentiment and stock returns. *Financial Analysts Journal* (March/April), 16-23.
- Fisher, K.L. and Statman, M. 2003. Consumer confidence and stock returns. *The Journal of Portfolio Management*, **30**(1), pp. 115-127.
- Fotak, V. 2007. The impact of blog recommendations on security prices and trading volumes. Working Paper, 1-42.
- Freitas AA. 2002. Data mining and knowledge discovery with evolutionary algorithms. Springer, Berlin, pp. 264.
- French, K.R. and Roll, R. 1986. Stock return variances: The arrival of information and the reaction of traders. *Journal of Financial Economics*, **17**(1), pp. 5-26.
- Friedman, M. 1953. The Case for Flexible Exchange Rates, *Essays in Positive Economics*. Chicago. University of Chicago Press.
- Friedman, N., Geiger, D. and Goldszmidt, M. 1997. Bayesian network classifiers. *Machine Learning*, **29**(2-3), pp. 131-163.
- Gallagher, L.A. and Taylor, M.P. 2002. Permanent and temporary components of stock prices: Evidence from assessing macroeconomic shocks. *Southern Economic Journal*, pp. 345-362.
- Gallant, A.R., Rossi, P.E. and Tauchen, G. 1992. Stock prices and volume. *Review of Financial Studies*, **5**(2), pp. 199-242.
- Gamon, M. 2004. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis, *Proceedings of the 20th international conference on Computational Linguistics 2004*, Association for Computational Linguistics, pp. 841.
- Gamon, M., Aue, A., Corston-Oliver, S. and Ringger, E. 2005. Pulse: Mining customer opinions from free text. *Advances in Intelligent Data Analysis VI*. Springer, pp. 121-132.
- Garfinkel, J.A. and Sokobin, J. 2006. Volume, opinion divergence, and returns: A study of post-earnings announcement drift. *Journal of Accounting Research*, pp. 85-112.

References

Garman, M.B. and Klass, M.J. 1980. On the estimation of security price volatilities from historical data. *Journal of business*, pp. 67-78.

Gervais, S. and Odean, T. 2001. Learning to be overconfident. *Review of Financial Studies*, **14**(1), pp. 1-27.

Giannini, R.C., Irvine, P.J. and Shu, T. 2014. Do local investors know more? A direct examination of individual investors' information set, *A Direct Examination of Individual Investors' Information Set (August 8, 2014)*. Asian Finance Association (AsFA) 2013 Conference 2014.

Giannini, R.C., Irvine, P.J. and Shu, T. 2013. The convergence and divergence of investors' opinions around earnings news: Evidence from a social network. Working Paper.

Gibbs, G. 2002. *Qualitative data analysis: Explorations with NVivo (Understanding social research)*. Buckingham: Open University Press.

Gidofalvi, G. and Elkan, C. 2001. Using news articles to predict stock price movements. *Department of Computer Science and Engineering, University of California, San Diego*.

Gilbert, E. and Karahalios, K. 2010. Widespread worry and the stock market, *Proceedings of the international conference on weblogs and social media 2010*, pp. 229-247.

Gilbert, N., 2008. *Researching social life*. London: Sage Publication.

Gilly, M.C. 1988. Sex roles in advertising: A comparison of television advertisements in Australia, Mexico, and the United States. *The Journal of marketing*, pp. 75-85.

Glosten, L.R., Jagannathan, R. and Runkle, D.E. 1993. On the relation between the expected value and the volatility of the nominal excess return on stocks. *The journal of finance*, **48**(5), pp. 1779-1801.

Glosten, L.R. and Milgrom, P.R. 1985. Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, **14**(1), pp. 71-100.

Goebel, R., Roebroek, A., Kim, D. and Formisano, E. 2003. Investigating directed cortical interactions in time-resolved fMRI data using vector autoregressive modeling and Granger causality mapping. *Magnetic resonance imaging*, **21**(10), pp. 1251-1261.

Goldbaum, M.H., Sample, P.A., Chan, K., Williams, J., Lee, T., Blumenthal, E., Girkin, C.A., Zangwill, L.M., Bowd, C. and Sejnowski, T. 2002. Comparing machine learning classifiers for diagnosing glaucoma from standard automated perimetry. *Investigative ophthalmology and visual science*, **43**(1), pp. 162-169.

Granger, C.W. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pp. 424-438.

References

- Gray, D.E. 2013. *Doing research in the real world*. London: Sage Publication.
- Green, T.C. 2006. The value of client access to analyst recommendations. *Journal of Financial and Quantitative Analysis*, **41**(01), pp. 1-24.
- Gromb, D. and Vayanos, D. 2002. Equilibrium and welfare in markets with financially constrained arbitrageurs. *Journal of Financial Economics*, **66**(2), pp. 361-407.
- Grossman, S. 1976. On the efficiency of competitive stock markets where trades have diverse information. *The Journal of finance*, **31**(2), pp. 573-585.
- Grossman, S.J. 1995. Dynamic asset allocation and the informational efficiency of markets. *The Journal of Finance*, **50**(3), pp. 773-787.
- Grossman, S.J. and Stiglitz, J.E. 1980. On the impossibility of informationally efficient markets. *The American Economic Review*, pp. 393-408.
- Gruhl, D., Guha, R., Kumar, R., Novak, J. and Tomkins, A. 2005. The predictive power of online chatter, *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining 2005*, ACM, pp. 78-87.
- Gu, B., Konana, P., Liu, A., Rajagopalan, B. and Ghosh, J. 2006. Identifying Information in Stock Message Boards and Its Implications for Stock Market Efficiency, *Workshop on Information Systems and Economics 2006*.
- Gu, B., Konana, P., Liu, A., Rajagopalan, B. and Ghosh, J. 2006. Predictive value of stock message board sentiments. *McCombs Research Paper No.IROM-11-06*.
- Guba, E.G. 1990. *The paradigm dialog*. London: Sage Publications.
- Guba, E. and Lincoln, Y., 1989. *Fourth generation evaluation*. Beverly Hills. CA: Sage.
- Gupta, V. and Lehal, G.S. 2009. A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence*, **1**(1), pp. 60-76.
- Guresen, E., Kayakutlu, G. and Daim, T.U. 2011. Using artificial neural network models in stock market index prediction. *Expert Systems with Applications*, **38**(8), pp. 10389-10397.
- Guyon, I. and Elisseeff, A. 2003. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, **3**, pp. 1157-1182.
- Hall, M., Frank, E., Holmes, G. Pfahringer, B., Reutemann, P. and Witten, I.H., 2009. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, **11**(1), pp. 10-18.
- Harris, M. and Raviv, A. 1993. Differences of opinion make a horse race. *Review of Financial Studies*, **6**(3), pp. 473-506.

References

- Harrison, J.M. and Kreps, D.M. 1978. Speculative investor behavior in a stock market with heterogeneous expectations. *The Quarterly Journal of Economics*, pp. 323-336.
- Hart, C. 1998. *Doing a literature review: Releasing the social science research imagination*. London:Sage Publication.
- Hiemstra, C. and Jones, J.D. 1994. Testing for linear and nonlinear Granger causality in the stock price-volume relation. *The Journal of Finance*, **49**(5), pp. 1639-1664.
- Hirshleifer, D. 2001. Investor psychology and asset pricing. *The Journal of Finance*, **56**(4), pp. 1533-1597.
- Hirshleifer, D. and Hong Teoh, S. 2003. Herd behaviour and cascading in capital markets: A review and synthesis. *European Financial Management*, **9**(1), pp. 25-66.
- Hirshleifer, J. 1977. The theory of speculation under alternative regimes of markets. *The Journal of Finance*, **32**(4), pp. 975-999.
- Hodrick, R.J. 1992. Dividend yields and expected stock returns: Alternative procedures for inference and measurement. *Review of Financial Studies*, **5**(3), pp. 357-386.
- Hong, H., Kubik, J.D. and Stein, J.C. 2005. Thy Neighbor's Portfolio: Word-of-Mouth Effects in the Holdings and Trades of Money Managers. *The Journal of Finance*, **60**(6), pp. 2801-2824.
- Hong, H., Lim, T. and Stein, J.C. 2000. Bad news travels slowly: Size, analyst coverage, and the profitability of momentum strategies. *The Journal of Finance*, **55**(1), pp. 265-295.
- Hong, H. and Stein, J.C. 2007. Disagreement and the Stock Market (Digest Summary). *Journal of Economic perspectives*, **21**(2), pp. 109-128.
- Hong, H. and Stein, J.C. 2003. Differences of opinion, short-sales constraints, and market crashes. *Review of Financial Studies*, **16**(2), pp. 487-525.
- 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.
- Houston, M.B. 2004. Assessing the validity of secondary data proxies for marketing constructs. *Journal of Business Research*, **57**(2), pp. 154-161.
- Hsieh, H.F. and Shannon, S.E. 2005. Three approaches to qualitative content analysis. *Qualitative health research*, **15**(9), pp. 1277-1288.
- Hsu, S., Hsieh, J.P., Chih, T. and Hsu, K. 2009. A two-stage architecture for stock price forecasting by integrating self-organizing map and support vector regression. *Expert Systems with Applications*, **36**(4), pp. 7947-7951.

References

- Huang, B. and Yang, C. 2001. An empirical investigation of trading volume and return volatility of the Taiwan Stock Market. *Global Finance Journal*, **12**(1), pp. 55-77.
- Huang, C. and Tsai, C. 2009. A hybrid SOFM-SVR with a filter-based feature selection for stock market forecasting. *Expert Systems with Applications*, **36**(2), pp. 1529-1539.
- Huang, C., Yang, D. and Chuang, Y. 2008. Application of wrapper approach and composite classifier to the stock trend prediction. *Expert Systems with Applications*, **34**(4), pp. 2870-2878.
- Huang, W., Nakamori, Y. and Wang, S. 2005. Forecasting stock market movement direction with support vector machine. *Computers & Operations Research*, **32**(10), pp. 2513-2522.
- Huang, D. and Chow, T.W. 2007. Identifying the biologically relevant gene categories based on gene expression and biological data: an example on prostate cancer. *Bioinformatics (Oxford, England)*, **23**(12), pp. 1503-1510.
- Huberman, G. 2001. Familiarity breeds investment. *Review of Financial Studies*, **14**(3), pp. 659-680.
- Hussey, J. and Hussey, R. 1997. Business research. A practical guide for undergraduate and postgraduate students. *Houndsmills: Macmillan*.
- Indurkha, N. and Zhang, T. 2005. *Text mining: predictive methods for analyzing unstructured information*, New York: Springer.
- Inza, I., Larrañaga, P., Blanco, R. and Cerrolaza, A.J. 2004. Filter versus wrapper gene selection approaches in DNA microarray domains. *Artificial Intelligence in Medicine*, **31**(2), pp. 91-103.
- Irvine, P.J. and Giannini, R.C. 2012. The Impact of Divergence of Opinions about Earnings within a Social Network. Working Paper.
- Jansen, D.W. and Tsai, C. 2010. Monetary policy and stock returns: Financing constraints and asymmetries in bull and bear markets. *Journal of Empirical finance*, **17**(5), pp. 981-990.
- Java, A., Song, X., Finin, T. and Tseng, B. 2007. Why we twitter: understanding microblogging usage and communities, *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis 2007*, ACM, pp. 56-65.
- Jegadeesh, N. and Titman, S. 2001. Profitability of momentum strategies: An evaluation of alternative explanations. *The Journal of Finance*, **56**(2), pp. 699-720.
- Jegadeesh, N. and Titman, S. 1993. Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, **48**(1), pp. 65-91.

References

- Jensen, M. 1978. Some anomalous evidence regarding market efficiency. *Journal of Financial Economics*, **6**(2/3), pp. 95-101.
- John, G.H., Kohavi, R. and Pflieger, K. 1994. Irrelevant Features and the Subset Selection Problem. *ICML 1994*, pp. 121-129.
- John, G.H., Kohavi, R. and Pflieger, K. 1994. Irrelevant Features and the Subset Selection Problem. *ICML 1994*, pp. 121-129.
- Johnson, B. and Christensen, L. 2000. Educational research: Quantitative and qualitative approaches.
- Johnson, B. and Turner, L.A. 2003. Data collection strategies in mixed methods research. *Handbook of mixed methods in social and behavioral research*, pp. 297-319.
- Johnson, P. and Harris, D. 2002. Qualitative and quantitative issues in research design. *Essential Skills for Management Research*, Sage, London, pp. 99-116.
- Johnson, R.B. and Onwuegbuzie, A.J. 2004. Mixed methods research: A research paradigm whose time has come. *Educational researcher*, **33**(7), pp. 14-26.
- Jones, A.L. 2006. Have Internet message boards changed market behavior? *Info*, **8**(5), pp. 67-76.
- Jones, S.S., Smith, L.B. and Landau, B. 1991. Object properties and knowledge in early lexical learning. *Child development*, **62**(3), pp. 499-516.
- Jorissen, R.N. and Gilson, M.K. 2005. Virtual screening of molecular databases using a support vector machine. *Journal of chemical information and modeling*, **45**(3), pp. 549-561.
- Kahneman, D. and Tversky, A. 1979. Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, pp. 263-291.
- Kamara, A., Miller, T.W. and Siegel, A.F. 1992. The effect of futures trading on the stability of Standard and Poor 500 returns. *Journal of Futures Markets*, **12**(6), pp. 645-658.
- Kandel, E. and Pearson, N.D. 1995. Differential interpretation of public signals and trade in speculative markets. *Journal of Political Economy*, pp. 831-872.
- Kaplan, B. and Maxwell, J.A. 2005. Qualitative research methods for evaluating computer information systems. *Evaluating the organizational impact of healthcare information systems*. Springer, pp. 30-55.
- Kaplanski, G. and Levy, H. 2010. Exploitable predictable irrationality: the FIFA World Cup effect on the US stock market. *Journal of Financial and Quantitative Analysis*, **45**(2), pp. 535.

References

- Kara, Y., Boyacioglu, M.A. and Baykan, O.K. 2011. Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange. *Expert Systems with Applications*, **38**(5), pp. 5311-5319.
- Karanikas, H. and Theodoulidis, B., 2002. Knowledge discovery in text and text mining software. *Centre for Research in Information Management, Department of Computation*.
- Karlsson, N., Loewenstein, G. and Seppi, D., 2009. The ostrich effect: Selective attention to information. *Journal of Risk and Uncertainty*, **38**(2), pp. 95-115.
- Kramer, A.D., 2010, April. An unobtrusive behavioral model of gross national happiness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 287-290. ACM.
- Karpoff, J.M. 1986. A theory of trading volume. *The Journal of Finance*, **41**(5), pp. 1069-1087.
- Kavussanos, M.G. and Dockery, E., 2001. A multivariate test for stock market efficiency: the case of ASE. *Applied Financial Economics*, **11**(5), pp. 573-579.
- Kelly, M. 1997. Do noise traders influence stock prices? *Journal of Money, Credit, and Banking*, pp. 351-363.
- Kim, K. 2003. Financial time series forecasting using support vector machines. *Neurocomputing*, **55**(1), pp. 307-319.
- Kim, O. and Verrecchia, R.E. 1991. Trading volume and price reactions to public announcements. *Journal of accounting research*, **29**(2), pp. 302-321.
- Kim, Y., Street, W.N. and Menczer, F. 2000. Feature selection in unsupervised learning via evolutionary search, *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining 2000*, ACM, pp. 365-369.
- Kira, K. and Rendell, L.A. 1992. The feature selection problem: Traditional methods and a new algorithm, *AAAI 1992*, pp. 129-134.
- Kling, G. and Gao, L. 2008. Chinese institutional investors' sentiment. *Journal of International Financial Markets, Institutions and Money*, **18**(4), pp. 374-387.
- Koenker, R. and Bassett Jr, G. 1978. Regression quantiles. *Econometrica: journal of the Econometric Society*, pp. 33-50.
- Koenker, R. and Hallock, K. 2001. Quantile regression: An introduction. *Journal of Economic Perspectives*, **15**(4), pp.43-56.
- Kohavi, R. and John, G.H. 1997. Wrappers for feature subset selection. *Artificial Intelligence*, **97**(1), pp. 273-324.

References

- Koski, J., Rice, E., and Tarhouni, A. 2004. Noise trading and volatility: Evidence from day trading and message boards. Working paper, University of Washington.
- Kothari, S., Li, X. and Short, J.E. 2009. The effect of disclosures by management, analysts, and business press on cost of capital, return volatility, and analyst forecasts: a study using content analysis. *The Accounting Review*, **84**(5), pp. 1639-1670.
- Kramer, A.D. 2010. An unobtrusive behavioral model of gross national happiness, *Proceedings of the 28th international conference on Human factors in computing systems 2010*, ACM, pp. 287-290.
- Krippendorff, K. 2012. *Content analysis: An introduction to its methodology*. London:Sage Publication.
- Kuechler, W. and Vaishnavi, V. 2008. The emergence of design research in information systems in North America. *Journal of Design Research*, **7**(1), pp. 1.
- Kukar, M. and Kononenko, I. 1998. Cost-Sensitive Learning with Neural Networks. *ECAI 1998*, pp. 445-449.
- Kumar, A. and Lee, C. 2006. Retail investor sentiment and return comovements. *The Journal of Finance*, **61**(5), pp. 2451-2486.
- Kumar, M. and Thenmozhi, M. 2006. Forecasting stock index movement: A comparison of support vector machines and random forest, *Indian Institute of Capital Markets 9th Capital Markets Conference Paper 2006*.
- Kurov, A. 2010. Investor sentiment and the stock market's reaction to monetary policy. *Journal of Banking and Finance*, **34**(1), pp. 139-149.
- Kurov, A., 2008. Investor sentiment, trading behavior and informational efficiency in index futures markets. *Financial Review*, **43**(1), pp. 107-127.
- Kyle, A.S. 1985. Continuous auctions and insider trading. *Econometrica: Journal of the Econometric Society*, pp. 1315-1335.
- Lagus, K., Honkela, T., Kaski, S., and Kohonen, T. (2000). WEBSOM for textual data mining. *Artificial Intelligence-Review*, 13:345–64.
- Lai, R.K., Fan, C., Huang, W. and Chang, P. 2009. Evolving and clustering fuzzy decision tree for financial time series data forecasting. *Expert Systems with Applications*, **36**(2), pp. 3761-3773.
- Lakonishok, J. and Maberly, E. 1990. The weekend effect: Trading patterns of individual and institutional investors. *The Journal of Finance*, **45**(1), pp. 231-243.
- Lakonishok, J., Shleifer, A. and Vishny, R.W. 1992. The impact of institutional trading on stock prices. *Journal of Financial Economics*, **32**(1), pp. 23-43.

References

- Lakonishok, J. and Smidt, S. 1988. Are seasonal anomalies real? A ninety-year perspective. *Review of Financial Studies*, **1**(4), pp. 403-425.
- Lamoureux, C.G. and Lastrapes, W.D. 1990. Heteroskedasticity in stock return data: volume versus GARCH effects. *The Journal of Finance*, **45**(1), pp. 221-229.
- Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D. and Allan, J. 2000. Language models for financial news recommendation, *Proceedings of the ninth international conference on Information and knowledge management 2000*, ACM, pp. 389-396.
- Lee, B. and Rui, O.M. 2002. The dynamic relationship between stock returns and trading volume: Domestic and cross-country evidence. *Journal of Banking and Finance*, **26**(1), pp. 51-78.
- Lee, C., Shleifer, A. and Thaler, R.H. 1991. Investor sentiment and the closed-end fund puzzle. *The Journal of Finance*, **46**(1), pp. 75-109.
- Lee, M. 2009. Using support vector machine with a hybrid feature selection method to the stock trend prediction. *Expert Systems with Applications*, **36**(8), pp. 10896-10904.
- Lee, W.Y., Jiang, C.X. and Indro, D.C. 2002. Stock market volatility, excess returns, and the role of investor sentiment. *Journal of Banking and Finance*, **26**(12), pp. 2277-2299.
- Lee, W.Y., Jiang, C.X. and Indro, D.C. 2002. Stock market volatility, excess returns, and the role of investor sentiment. *Journal of Banking and Finance*, **26**(12), pp. 2277-2299.
- Lehmann, D.R. 1989. *Market research and analysis*, Homewood, IL: Irwin.
- Lemon, J., Degenhardt, L., Slade, T. and Mills, K. 2010. Quantitative Data Analysis. *Addiction Research Methods*, pp. 163-183.
- Lerman, A. 2010. *Individual investors' attention to accounting information: Message board discussions*. Working Paper, Yale University.
- Levis, M. 1989. Stock market anomalies: A re-assessment based on the UK evidence. *Journal of Banking and Finance*, **13**(4), pp. 675-696.
- Li, X., Rao, S., Wang, Y. and Gong, B. 2004. Gene mining: a novel and powerful ensemble decision approach to hunting for disease genes using microarray expression profiling. *Nucleic acids research*, **32**(9), pp. 2685-2694.
- Li, Y., Xie, M. and Goh, T. 2009. A study of mutual information based feature selection for case based reasoning in software cost estimation. *Expert Systems with Applications*, **36**(3), pp. 5921-5931.
- Lin, C., Yeh, C., Liang, S., Chung, J. and Kumar, N. 2006. Support-vector-based fuzzy neural network for pattern classification. *Fuzzy Systems, IEEE Transactions on*, **14**(1), pp. 31-41.

References

- Lin, J., Cheng, C. and Chau, K. 2006. Using support vector machines for long-term discharge prediction. *Hydrological Sciences Journal*, **51**(4), pp. 599-612.
- Lincoln, Y.S. and Guba, E.G. 1985. *Naturalistic inquiry*. Beverly Hills, CA; London: Sage Publication.
- Lindzon, H., Pearlman, P., and Ivanhoff, I. 2011. *The StockTwits Edge*. New Jersey: John Wiley and Sons.
- Liu, B., Hu, M. and Cheng, J. 2005. Opinion observer: analyzing and comparing opinions on the web, *Proceedings of the 14th international conference on World Wide Web 2005*, ACM, pp. 342-351.
- Liu, H., Motoda, H. and Yu, L. 2002. Feature selection with selective sampling, *ICML 2002*, pp. 395-402.
- Liu, H. and Yu, L. 2005. Toward integrating feature selection algorithms for classification and clustering. *Knowledge and Data Engineering, IEEE Transactions on*, **17**(4), pp. 491-502.
- Liu, S. 2015. Investor sentiment and stock market liquidity. *Journal of Behavioral Finance*, **16**(1), pp. 51-67.
- Lo, A.W. and Mackinlay, A.C. 2011. *A non-random walk down Wall Street*. Princeton University Press.
- Lo, A.W. and Mackinlay, A.C. 1988. Stock market prices do not follow random walks: Evidence from a simple specification test. *Review of Financial Studies*, **1**(1), pp. 41-66.
- Long, J.B., Shleifer, A., Summers, L.H. and Waldmann, R.J. 1990. Positive feedback investment strategies and destabilizing rational speculation. *The Journal of Finance*, **45**(2), pp. 379-395.
- Loughran, T. and McDonald, B. 2011. When is a liability not a liability? Textual analysis, dictionaries and 10-Ks. *The Journal of Finance*, **66**(1), pp. 35-65.
- Lu, J., Plataniotis, K.N. and Venetsanopoulos, A.N. 2003. Face recognition using kernel direct discriminant analysis algorithms. *Neural Networks, IEEE Transactions on*, **14**(1), pp. 117-126.
- Lucey, B.M. 2005. Does volume provide information? Evidence from the Irish stock market. *Applied Financial Economics Letters*, **1**(2), pp. 105-109.
- Lutkepohl, H. 2005. New introduction to multiple time series. *Springer Verlag, Berlin*, **2**, pp. 70-78.
- Ma, L. and Pohlman, L. 2008. Return forecasts and optimal portfolio construction: a quantile regression approach. *The European Journal of Finance*, **14**(5), pp. 409-425.

References

- Malkiel, B.G. 2003. The efficient market hypothesis and its critics. *The Journal of Economic Perspectives*, **17**(1), pp. 59-82.
- Malkiel, B. G. 1973. *A Random Walk Down Wall Street*. W.W. Norton, New York.
- Mao, H., Counts, S. and Bollen, J. 2011. Predicting financial markets: Comparing survey, news, twitter and search engine data. *arXiv preprint arXiv:1112.1051*.
- Mao, Y., Wei, W., Wang, B. and Liu, B. 2012. Correlating S&P 500 stocks with Twitter data, *Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research 2012*, ACM, pp. 69-72.
- Mark, H., Eibe, F., Geoffrey, H., Bernhard, P., Peter, R., Ian, H.W. 2009. The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.
- Maxwell, S.E. and Delaney, H.D. 2004. *Designing experiments and analyzing data: A model comparison perspective*. Psychology Press.
- Maylor, H. and Blackmon, K., 2005. *Researching business and management: a roadmap for success*. New York: Palgrave Macmillan.
- Mckee, A., 2003. *Textual analysis: A beginner's guide*. London: Sage Publication.
- Mcmillan, D.G. 2005. Non-linear dynamics in international stock market returns. *Review of Financial Economics*, **14**(1), pp. 81-91.
- Miles, M.B. and Huberman, A.M. 1994. *Qualitative data analysis: An expanded sourcebook*. London:Sage Publications.
- Milgrom, P. and Stokey, N. 1982. Information, trade and common knowledge. *Journal of Economic Theory*, **26**(1), pp. 17-27.
- Miller, A. 2002. Subset selection in regression. *Monographs on statistics and applied probability (95) Show all parts in this series*.
- Miller, E.M. 1977. Risk, uncertainty, and divergence of opinion. *The Journal of Finance*, **32**(4), pp. 1151-1168.
- Mingers, J. 2001. Combining IS research methods: towards a pluralist methodology. *Information systems research*, **12**(3), pp. 240-259.
- Mishne, G. and Glance, N. 2006. Predicting movie sales from blogger sentiment, *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs 2006*, pp. 11.
- Mitchell, M.L. and Mulherin, J.H., 1994. The impact of public information on the stock market. *The Journal of Finance*, **49**(3), pp. 923-950.
- Mitchell, T.M. 1997. Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, **45**.

References

- Mittermayer, M. 2004. Forecasting intraday stock price trends with text mining techniques, *System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on 2004*, IEEE, pp. 10 pp.
- Mizrach, B. and Weerts, S. 2009. Experts online: An analysis of trading activity in a public Internet chat room. *Journal of Economic Behavior and Organization*, **70**(1), pp. 266-281.
- Mukherjee, S., Osuna, E. and Girosi, F. 1997. Nonlinear prediction of chaotic time series using support vector machines, *Neural Networks for Signal Processing [1997] VII. Proceedings of the 1997 IEEE Workshop 1997*, IEEE, pp. 511-520.
- Muller, K., Smola, A.J., Ratsch, G., Scholkopf, B., Kohlmorgen, J. and Vapnik, V. 1997. Predicting time series with support vector machines. *Artificial Neural Networks—ICANN'97*. Springer, pp. 999-1004.
- Myers, M.D. 2013. *Qualitative research in business and management*. London: Sage Publication.
- Myers, M.D. 1997. Qualitative research in information systems. *Management Information Systems Quarterly*, **21**, pp. 241-242.
- Myers, M.D. and Avison, D. 2002. *Qualitative research in information systems*. Sage Publication, London.
- Nagel, T., 1989. *The view from nowhere*. Oxford university press.
- Neal, R. and Wheatley, S.M. 1998. Do measures of investor sentiment predict returns? *Journal of Financial and Quantitative Analysis*, **33**(4).
- Nelson, D.B., 1991. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica: Journal of the Econometric Society*, pp. 347-370.
- Neuman, W.L., 2005. *Social research methods: Quantitative and qualitative approaches*. Allyn and Bacon Boston.
- Ng, L. and Wu, F., 2006. Peer effects in investor trading decisions: evidence from a natural experiment, *Western Finance Association meeting 2006*.
- Ni, L., Ni, Z. and Gao, Y. 2011. Stock trend prediction based on fractal feature selection and support vector machine. *Expert Systems with Applications*, **38**(5), pp. 5569-5576.
- Nofsinger, J.R. 2005. Social mood and financial economics. *The Journal of Behavioral Finance*, **6**(3), pp. 144-160.
- O'connor, B., Balasubramanian, R., Routledge, B.R. and Smith, N.A., 2010. From tweets to polls: Linking text sentiment to public opinion time series, *Proceedings of the International AAAI Conference on Weblogs and Social Media 2010*, pp. 122-129.

References

Oates, B.J. 2006. *Researching Information Systems and Computing*. London: Sage Publication.

Odean, T. 1998a. Are investors reluctant to realize their losses? *The Journal of finance*, **53**(5), pp. 1775-1798.

Odean, T. 1998b. Volume, volatility, price, and profit when all traders are above average. *The Journal of Finance*, **53**(6), pp. 1887-1934.

Odean, T. 1999. Do Investors Trade Too Much? *The American Economic Review*, **89**(5), pp. 1279-1298.

Oh, C. and Sheng, O. (2011), "Investigating Predictive Power of Stock Micro Blog Sentiment in Forecasting Future Stock Price Directional Movement", in 32nd *International Conference on Information Systems*, AIS, p.17.

Oliveira, N., Cortez, P. and Areal, N. 2014. Automatic creation of stock market lexicons for sentiment analysis using StockTwits data, *Proceedings of the 18th International Database Engineering and Applications Symposium 2014*, ACM, pp. 115-123.

Oliveira, N., Cortez, P. and Areal, N. 2013. On the Predictability of Stock Market Behavior Using StockTwits Sentiment and Posting Volume. *Progress in Artificial Intelligence*. Springer, pp. 355-365.

O'reilly, T. 2007. What is Web 2.0: Design patterns and business models for the next generation of software. *Communications and strategies*, (1), pp. 17.

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.

Pang, B. and Lee, L. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, **2**(1-2), pp. 1-135.

Pang, B. and Lee, L. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales, *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics 2005*, Association for Computational Linguistics, pp. 115-124.

Pang, B., Lee, L. and Vaithyanathan, S. 2002. Thumbs up?: sentiment classification using machine learning techniques, *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 2002*, Association for Computational Linguistics, pp. 79-86.

Pearl, J. 1988. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.

Person, K. 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, **2**(11), pp.559-572.

References

- Pfleiderer, P. 1984. *The volume of trade and the variability of prices: A framework for analysis in noisy rational expectations equilibria*. Graduate School of Business, Stanford University.
- Phillips, E. and Pugh, D. 2005. *How to Get a PhD: A Handbook for Students and their Supervisors*. England, Open University Press.
- Pinto, M.V. and Asnani, K. 2011. Stock price prediction using quotes and financial news. *International Journal of Soft Computing and Engineering (IJSCE)*, **1**(5).
- Polat, K. and Guneş, S. 2007. Breast cancer diagnosis using least square support vector machine. *Digital Signal Processing*, **17**(4), pp. 694-701.
- Popper, K.R. 1959. *The logic of scientific discovery*. London: Hutchinson.
- Porter, M.F. 1980. An algorithm for suffix stripping. *Program: electronic library and information systems*, **14**(3), pp. 130-137.
- Pozzebon, M. 2004. Conducting and evaluating critical interpretive research: examining criteria as a key component in building a research tradition. *Information Systems Research*. Springer, pp. 275-292.
- Prati, R.C., Batista, G.E. and Monard, M.C. 2004. Class imbalances versus class overlapping: an analysis of a learning system behavior. *MICAI 2004: Advances in Artificial Intelligence*. Springer, pp. 312-321.
- Pyun, C.S., Lee, S.Y. and Nam, K. 2001. Volatility and information flows in emerging equity market: A case of the Korean stock exchange. *International Review of Financial Analysis*, **9**(4), pp. 405-420.
- Qian, B. and Rasheed, K. 2007. Stock market prediction with multiple classifiers. *Applied Intelligence*, **26**(1), pp. 25-33.
- Qiu, G., Liu, B., Bu, J. and Chen, C. 2011. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, **37**(1), pp. 9-27.
- Quinlan, J. R. 1993. *C4.5: Programs for Machine Learning*. San Francisco, CA, Morgan Kaufmann.
- R Development Core Team. 2012. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL [http:// www.R-project.org/](http://www.R-project.org/).
- Ragunathan, V. and Peker, A. 1997. Price variability, trading volume and market depth: evidence from the Australian futures market. *Applied Financial Economics*, **7**(5), pp. 447-454.
- Ranco, G., Aleksovski, D., Caldarelli, G., Grcar, M. and Mozetic, I. 2015. The effects of Twitter sentiment on stock price returns. *PloS one*, **10**(9), p.e0138441.

References

- Rao, T. and Srivastava, S., 2014. Twitter sentiment analysis: How to hedge your bets in the stock markets. *State of the Art Applications of Social Network Analysis*. Springer, pp. 227-247.
- Reilly, F. K., Brown, K. C. 2009. "Efficient Capital Markets", in *Equity and Fixed Income*, CFA Program Curriculum, vol. 5, Pearson Custom Publishing.
- Rish, I. 2001. An empirical study of the Naive Bayes classifier, *IJCAI 2001 workshop on empirical methods in artificial intelligence 2001*, pp. 41-46.
- Ritterman, J., Osborne, M. and Klein, E. 2009. Using prediction markets and Twitter to predict a swine flu pandemic, *1st international workshop on mining social media 2009*.
- Robnik-Sikonja, M. and Kononenko, I. 2003. Theoretical and empirical analysis of ReliefF and RReliefF. *Machine Learning*, **53**(1-2), pp. 23-69.
- Rokach, L. and Maimon, O. 2005. Decision trees. *Data Mining and Knowledge Discovery Handbook*. Springer, pp. 165-192.
- Rudd, A.T. 1979. The Revised Dow Jones Industrial Average: New Wine in Old Bottles? *Financial Analysts Journal*, pp. 57-63.
- Ruiz, E.J., Hristidis, V., Castillo, C., Gionis, A. and Jaimes, A. 2012. Correlating financial time series with micro-blogging activity, *Proceedings of the fifth ACM international conference on Web search and data mining 2012*, ACM, pp. 513-522.
- Ruiz, R., Riquelme, J.C. and Aguilar-Ruiz, J.S. 2006. Incremental wrapper-based gene selection from microarray data for cancer classification. *Pattern Recognition*, **39**(12), pp. 2383-2392.
- Sabherwal, S., Sarkar, S.K. and Zhang, Y. 2008. Online talk: does it matter? *Managerial Finance*, **34**(6), pp. 423-436.
- Saeys, Y., Inza, I. and Larranaga, P. 2007. A review of feature selection techniques in bioinformatics. *bioinformatics*, **23**(19), pp. 2507-2517.
- Samuelson, P.A. 1965. Proof that properly anticipated prices fluctuate randomly. *Industrial management review*, **6**(2), pp. 41-49.
- Sarantakos, S. 2005. Social Research. 3rd. *Hampshire: Palgrave Macmillan* .
- Saunders, C., Gammerman, A. and Vovk, V. 1998. Ridge regression learning algorithm in dual variables, (*ICML-1998*) *Proceedings of the 15th International Conference on Machine Learning 1998*, Morgan Kaufmann, pp. 515-521.
- Saunders, M.N., Saunders, M., Lewis, P. and Thornhill, A. 2011. *Research methods for business students, 5/e*. Pearson Education India.

References

- Saxton, G.D. 2012. New Media and External Accounting Information: A Critical Review. *Australian Accounting Review*, **22**(3), pp. 286-302.
- Schmeling, M. 2009. Investor sentiment and stock returns: some international evidence. *Journal of Empirical Finance*, **16**(3), pp. 394-408.
- Schoelkopf and A. Smola.2002. *Learning with Kernels*. MIT Press, Cambridge MA.
- Schrag, F. 1992. In defense of positivist research paradigms. *Educational Researcher*, **21**(5), pp. 5-8.
- Schulhofer-Wohl, S. 2008. Heterogeneous risk preferences and the welfare cost of business cycles. *Review of Economic Dynamics*, **11**(4), pp. 761-780.
- Schumaker, R.P. and Chen, H. 2011. Predicting Stock Price Movement from Financial News Articles. *Information Systems for Global Financial Markets: Emerging Developments and Effects: Emerging Developments and Effects*, pp. 96.
- Schumaker, R.P. and Chen, H. 2009. Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions on Information Systems (TOIS)*, **27**(2), pp. 12.
- Schumaker, R.P., Zhang, Y., Huang, C. and Chen, H. 2012. Evaluating sentiment in financial news articles. *Decision Support Systems*, **53**(3), pp. 458-464.
- Schwert, G.W. 2003. Anomalies and market efficiency. *Handbook of the Economics of Finance*, **1**, pp. 939-974.
- Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, **34**(1), pp. 1-47.
- Sekaran, U., *Research Methods for Business: A Skill Building Approach*. 2003. *John Willey and Sons, New York*.
- Shalen, C.T. 1993. Volume, volatility, and the dispersion of beliefs. *Review of Financial Studies*, **6**(2), pp. 405-434.
- Shamoo, A.E. and Resnik, D.B. 2003. *Responsible conduct of research*. Oxford University Press.
- Sharif, A.M. 2004. *Knowledge representation within information systems in manufacturing environments*. Unpublished Thesis, Department of Information Systems and Computing, Brunel University, UK.
- Shiller, R.J., Fischer, S. and Friedman, B.M. 1984. Stock prices and social dynamics. *Brookings papers on economic activity*, pp. 457-510.
- Shiller, R.J., Kon-Ya, F. and Tsutsui, Y. 1996. Why did the Nikkei crash? Expanding the scope of expectations data collection. *The review of economics and statistics*, pp. 156-164.

References

Shleifer, A. 2000. *Inefficient markets: An introduction to behavioral finance*. Oxford University Press.

Shleifer, A. and Summers, L.H., 1990. The noise trader approach to finance. *The Journal of Economic Perspectives*, pp. 19-33.

Shleifer, A. and Vishny, R.W. 1997. The limits of arbitrage. *The Journal of Finance*, **52**(1), pp. 35-55.

Shmueli, G. and Koppius, O. 2010. Predictive analytics in information systems research. *Robert H. Smith School Research Paper No. RHS*, pp. 06-138.

Sias, R. 1997. The sensitivity of individual and institutional investors' expectations to changing market conditions: Evidence from closed-end funds. *Review of Quantitative Finance and Accounting*, **8**(3), pp. 245-269.

Sima, C. and Dougherty, E.R. 2008. The peaking phenomenon in the presence of feature-selection. *Pattern Recognition Letters*, **29**(11), pp. 1667-1674.

Simon, D.P. and Wiggins, R.A. 2001. S&P futures returns and contrary sentiment indicators. *Journal of Futures Markets*, **21**(5), pp. 447-462.

Singh, N., Hu, C. and Roehl, W.S. 2007. Text mining a decade of progress in hospitality human resource management research: identifying emerging thematic development. *International Journal of Hospitality Management*, **26**(1), pp. 131-147.

Smith, J.K. 1984. The problem of criteria for judging interpretive inquiry. *Educational evaluation and policy analysis*, pp. 379-391.

Smith, J.K. 1983. Quantitative versus qualitative research: An attempt to clarify the issue. *Educational researcher*, pp. 6-13.

Smola, A.J. and Scholkopf, B. 2004. A tutorial on support vector regression. *Statistics and computing*, **14**(3), pp. 199-222.

Snyder, B. and Barzilay, R. 2007. Multiple aspect ranking using the good grief algorithm, *Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference (HLT-NAACL) 2007*, pp. 300-307.

Solt, M.E. and Statman, M. 1988. How useful is the sentiment index?. *Financial Analysts Journal*, **44**(5), pp.45-55.

Sorensen, H.T., Sabroe, S. and Olsen, J. 1996. A framework for evaluation of secondary data sources for epidemiological research. *International journal of epidemiology*, **25**(2), pp. 435-442.

Spackman, K.A. 1989. Signal detection theory: Valuable tools for evaluating inductive learning, *Proceedings of the sixth international workshop on Machine learning 1989*, Morgan Kaufmann Publishers Inc., pp. 160-163.

References

- Spinakis, A. and Peristera, P. 2004. Text Mining Tools: Evaluation Methods and Criteria. *Text Mining and its Applications*. Springer, pp. 131-149.
- Sprenger, T.O., Tumasjan, A., Sandner, P.G. and Welpe, I.M. 2014. Tweets and trades: The information content of stock microblogs. *European Financial Management*, **20**(5), pp. 926-957.
- Stein, G., Chen, B., Wu, A.S. and Hua, K.A. 2005. Decision tree classifier for network intrusion detection with GA-based feature selection, *Proceedings of the 43rd annual Southeast regional conference-Volume 2 2005*, ACM, pp. 136-141.
- Stitson, M., Weston, J., Gammernan, A., Vovk, V. and Vapnik, V. 1996. Theory of support vector machines. *University of London*.
- StockTwits. 2008. About StockTwits. [ONLINE] Available at: <http://stocktwits.com/about>. [Accessed 18 September 13].
- Strapparava, C. and Mihalcea, R. 2008. Learning to identify emotions in text, *Proceedings of the 2008 ACM symposium on Applied computing 2008*, ACM, pp. 1556-1560.
- Strong, N. 1992. Modelling abnormal returns: a review article. *Journal of Business Finance and Accounting*, **19**(4), pp. 533-553.
- Su, F. and Markert, K. 2008. From words to senses: a case study of subjectivity recognition, *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1 2008*, Association for Computational Linguistics, pp. 825-832.
- Swets, J.A., Dawes, R.M. and Monahan, J. 2000. Better DECISIONS through. *Scientific American*, pp. 83.
- Tan, A. 1999. Text mining: The state of the art and the challenges, *Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases 1999*, pp. 65-70.
- Tan, A.C. and Gilbert, D. 2003. An empirical comparison of supervised machine learning techniques in bioinformatics, *Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics 2003-Volume 19 2003*, Australian Computer Society, Inc., pp. 219-222.
- Tashakkori, A. and Teddlie, C., 1998. *Mixed methodology: Combining qualitative and quantitative approaches*. London: Sage Publication.
- Teddlie, C. and Yu, F., 2007. Mixed methods sampling a typology with examples. *Journal of mixed methods research*, **1**(1), pp. 77-100.
- Tetlock, P.C. 2007. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, **62**(3), pp. 1139-1168.

References

- Tetlock, P.C., Saar-Tsechansky, M. and Macskassy, S. 2008. More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, **63**(3), pp. 1437-1467.
- Thaler, R. 1987. Anomalies: seasonal movements in security prices ii: weekend, holiday, turn of the month, and intraday effects. *The Journal of Economic Perspectives*, **1**(2), pp. 169-177.
- Thelwall, M., Buckley, K., Paltoglou, G., Cai, D. and Kappas, A. 2010. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, **61**(12), pp. 2544-2558.
- Thissen, U., Van Brakel, R., De Weijer, A., Melssen, W. and Buydens, L. 2003. Using support vector machines for time series prediction. *Chemometrics and Intelligent Laboratory Systems*, **69**(1), pp. 35-49.
- Toda, H.Y. and Yamamoto, T. 1995. Statistical inference in vector autoregressions with possibly integrated processes. *Journal of Econometrics*, **66**(1), pp. 225-250.
- Tomarken, A.J., 1995. A psychometric perspective on psychophysiological measures. *Psychological assessment*, **7**(3), pp. 387.
- Tong, S. and Koller, D. 2002. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, **2**, pp. 45-66.
- Tripathi, V. (2009) "Company Fundamentals and Equity Returns in India", *International Research Journal of Finance and Economics*, **29**, pp. 188-226.
- Trueman, B. and Titman, S. 1988. An explanation for accounting income smoothing. *Journal of accounting research*, pp. 127-139.
- Tumarkin, R. and Whitelaw, R.F., 2001. News or noise? Internet postings and stock prices. *Financial Analysts Journal*, pp. 41-51.
- Turney, P.D. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews, *Proceedings of the 40th annual meeting on association for computational linguistics 2002*, Association for Computational Linguistics, pp. 417-424.
- Ur-Rahman, N., 2010. *Textual Data Mining Applications for Industrial Knowledge Management Solutions*, PhD Thesis, Loughborough University, UK.
- Van Bommel, J. 2003, "Rumors", *The Journal of Finance*, vol. 58, no. 4, pp. 1499-1519.
- Vapnik, V. and Lerner, A. 1963. Generalized portrait method for pattern recognition. *Automation and Remote Control*, **24**(6), pp. 774-780.
- Vapnik, V. 1998. The support vector method of function estimation. *Nonlinear Modeling*. Springer, pp. 55-85.

References

- Vapnik, V. 1963. Pattern recognition using generalized portrait method. *Automation and remote control*, **24**, pp. 774-780.
- Vapnik, V.N. 2000. The nature of statistical learning theory. Statistics for Engineering and Information Science. *Springer-Verlag, New York*.
- Vapnik, V. and Chervonenkis, A. 1964. A note on one class of perceptrons. *Automation and Remote Control*, **25**(1).
- Varian, H. 1989. Differences of opinion in financial markets. In: Stone, C. (Ed.), *Financial Risk: Theory, Evidence and Implications*. Kluwer Academic Publications, Boston.
- Varian, H.R. 1985. Divergence of opinion in complete markets: A note. *The Journal of Finance*, **40**(1), pp. 309-317.
- Verma, R. and Verma, P. 2007. Noise trading and stock market volatility. *Journal of Multinational Financial Management*, **17**(3), pp. 231-243.
- Veropoulos, K., Campbell, C. and Cristianini, N. 1999. Controlling the sensitivity of support vector machines, *Proceedings of the international joint conference on artificial intelligence 1999*, Citeseer, pp. 55-60.
- Vinciotti, V., Tucker, A., Kellam, P. and Liu, X. 2006. Robust selection of predictive genes via a simple classifier. *Applied bioinformatics*, **5**(1), pp. 1-11.
- Vogt, W. 1993. *Dictionary of statistics and methodology*, London: Sage Publication.
- Voss, C., Tsikriktsis, N. and Frohlich, M. 2002. Case research in operations management. *International journal of operations and production management*, **22**(2), pp. 195-219.
- Walsham, G. 1995. Interpretive case studies in IS research: nature and method. *European Journal of information systems*, **4**(2), pp. 74-81.
- Wang, C. 2001. Investor sentiment and return predictability in agricultural futures markets. *Journal of Futures Markets*, **21**(10), pp. 929-952.
- Wang, F.A. 2001. Overconfidence, investor sentiment, and evolution. *Journal of Financial Intermediation*, **10**(2), pp. 138-170.
- Wang, G., Wang, T., Wang, B., Sambasivan, D., Zhang, Z., Zheng, H. and Zhao, B.Y., 2014. Crowds on wall street: Extracting value from social investing platforms. *arXiv preprint arXiv:1406.1137*, .
- Weber, R.P. 1990. *Basic content analysis*. Beverly Hills, CA; Sage Publication.
- Weiss, S.M., Indurkha, N., Zhang, T. and Damerau, F. 2010. *Text mining: predictive methods for analyzing unstructured information*. Springer Science and Business Media.

References

- Wermers, R. 1999. Mutual fund herding and the impact on stock prices. *The Journal of Finance*, **54**(2), pp. 581-622.
- Weston, J.F., Siu, J.A. and Johnson, B.A. 2001. *Takeovers, restructuring, and corporate governance*. Prentice Hall.
- White, K.J. and Sutcliffe, R.F. 2006. Applying incremental tree induction to retrieval from manuals and medical texts. *Journal of the American Society for Information Science and Technology*, **57**(5), pp. 588-600.
- Williams, G.J. 2009. Rattle: a data mining GUI for R. *The R Journal*, **1**(2), pp. 45-55.
- Wilson, T., Wiebe, J. and Hoffmann, P. 2005. Recognizing contextual polarity in phrase-level sentiment analysis, *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing 2005*, Association for Computational Linguistics, pp. 347-354.
- Wilson, T., Wiebe, J. and Hwa, R. 2006. Recognizing strong and weak opinion clauses. *Computational Intelligence*, **22**(2), pp. 73-99.
- Windle, P.E. 2010. Secondary Data Analysis: Is It Useful and Valid? *Journal of PeriAnesthesia Nursing*, **25**(5), pp. 322-324.
- Winer, R.S. 1999. Experimentation in the 21st century: the importance of external validity. *Journal of the Academy of Marketing Science*, **27**(3), pp. 349-358.
- Witten, I.H., Frank, E. and Hall, M.A. 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco.
- Witten, I.H. 2013. Data Mining with Weka: Training and Testing. Online course, University of Waikato, New Zealand.
- Witten, I.H., Frank, E., Trigg, L.E., Hall, M.A., Holmes, G. and Cunningham, S.J., 1999. *Weka: Practical machine learning tools and techniques with Java implementations*. Morgan Kaufmann, San Francisco.
- Wordle-Beautiful Word clouds-May 20 2014-<http://www.wordle.net/creat>.
- Wu, M., Lin, S. and Lin, C. 2006. An effective application of decision tree to stock trading. *Expert Systems with Applications*, **31**(2), pp. 270-274.
- Wysocki, P. 1998. Cheap talk on the web: The determinants of postings on stock message boards. *University of Michigan Business School Working Paper*, (98025).
- Xing, E.P., Jordan, M.I. and Karp, R.M. 2001. Feature selection for high-dimensional genomic microarray data, *ICML 2001*, pp. 601-608.
- Xu, F., 2012. Data Mining in Social Media for Stock Market Prediction. Master Thesis of Electronic Commerce at Dalhousie University Halifax, Nova Scotia.

References

- Xue-Shen, S., Zhong-Ying, Q., Da-Ren, Y., Qing-Hua, H. and Hui, Z. 2007. A novel feature selection approach using classification complexity for SVM of stock market trend prediction, *Management Science and Engineering, 2007. ICMSE 2007. International Conference on 2007*, IEEE, pp. 1654-1659.
- Yang, H., Chan, L. and King, I. 2002. Support vector machine regression for volatile stock market prediction. *Intelligent Data Engineering and Automated Learning—IDEAL 2002*. Springer, pp. 391-396.
- Yang, J. and Olafsson, S. 2006. Optimization-based feature selection with adaptive instance sampling. *Computers and Operations Research*, **33**(11), pp. 3088-3106.
- Yang, Y. and Pedersen, J.O. 1997. A comparative study on feature selection in text categorization, *ICML 1997*, pp. 412-420.
- Yin, R.K. 2013. *Case study research: Design and methods*. London: Sage Publications.
- Yu, J. and Yuan, Y. 2011. Investor sentiment and the mean–variance relation. *Journal of Financial Economics*, **100**(2), pp. 367-381.
- Yu, L. and Liu, H. 2004. Efficient feature selection via analysis of relevance and redundancy. *The Journal of Machine Learning Research*, **5**, pp. 1205-1224.
- Yuan, Y. 2008. Attention and trading. *Unpublished Working Paper. University of Pennsylvania*.
- Zhang, C., Zeng, D., Li, J., Wang, F. and Zuo, W. 2009. Sentiment analysis of Chinese documents: From sentence to document level. *Journal of the American Society for Information Science and Technology*, **60**(12), pp. 2474-2487.
- Zhang, D., Chen, S. and Zhou, Z. 2008. Constraint Score: A new filter method for feature selection with pairwise constraints. *Pattern Recognition*, **41**(5), pp. 1440-1451.
- Zhang, X., Hu, Y., Xie, K., Wang, S., Ngai, E. and Liu, M. 2014. A causal feature selection algorithm for stock prediction modeling. *Neurocomputing*, **142**, pp. 48-59.
- Zhang, X., Fuehres, H. and Gloor, P.A. 2011. Predicting stock market indicators through twitter “I hope it is not as bad as I fear”. *Procedia-Social and Behavioral Sciences*, **26**, pp. 55-62.
- Zhang, Y. and Wildemuth, B.M. 2009. Qualitative analysis of content. *Applications of social research methods to questions in information and library science*, pp. 308-319.
- Zhang, Y., Luo, A. and Zhao, Y. 2004. An automated classification algorithm for multiwavelength data, *Proceedings of SPIE 2004*, pp. 483-490.
- Zhang, Y. and Swanson, P.E. 2010. Are day traders bias free? Evidence from Internet stock message boards. *Journal of Economics and Finance*, **34**(1), pp. 96-112.

References

Zhao, Y., Karypis, G. and Fayyad, U. 2005. Hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery*, **10**(2), pp. 141-168.

Zheng, H. and Zhang, Y. 2008. Feature selection for high-dimensional data in astronomy. *Advances in Space Research*, **41**(12), pp. 1960-1964.

Zikmund, W.G. 2000. *Business research methods*. 6th ed. Fort Worth, TX; London: Dryden Press, Harcourt Brace College Publishers.

APPENDICES

Appendix I

Inter coder Agreement

Agreement Reliability and Validity

Reliability and validity are the two criteria used to evaluate the quality of research involving inter-coder agreement. Reliability is the extent to which agreement yields the same result if replicated over time and is said to be consistent. The fundamental assumption behind the methodology discussed in this thesis is that the data can be considered reliable if coders can be shown to agree on the coding process of classifying and assigning tweet postings into the three distinct classes (Craggs et al., 2005). If both coders have produced consistently similar results, it can be concluded that both have gained a similar understanding of the coding rules and categories, and greater consistency in the coding process has been achieved. Without the agreement of independent coders able to repeat the coding process and produce consistent results, researchers would be unable to draw satisfactory conclusions to confirm theories or make inferences about the generalisability of their research.

On the other hand, validity refers to the degree to which a study accurately assesses and measures what it is assumed to be measuring (David and Sutton 2004, p.171). Burns and Bush (1995) argue that the validity of a measurement instrument refers to how well it captures what a researcher has designed and set out to measure. In the context of this research study, the validity of the coding scheme is of critical importance for showing that the coding scheme captures the “truth” of the phenomenon under study (Artstein and Poesio, 2008). The validity of the coding scheme was tested by the following means: (1) The researcher determined that the categorical variables and the coding scheme adopted for the study have been defined and used previously in the literature (Churchill and Iacobucci, 2009). The advantage of adopting a coding scheme developed in previous studies is that it allows a comparison of research findings across several studies (Artstein and Poesio, 2008). (2) The researcher defined and identified the coding rules, verifying that they are clearly stated in the codebook and that the categorical variables are well defined and established. (3) The researcher consistently revised the coding scheme while carrying out the classification process to ensure regular updating of the coding scheme.

Appendices

(4) Critical discussions between coders and re-evaluations of their classification results were carried out to ensure a high level of consistency between the two independent coders.

Inter-coder Agreement Results

As with most text classification techniques using a hand-labelled training set (e.g., Antweiler and Frank (2004b); Sprenger et al., (2014)), classification or a review by a second judge may be required to ensure accurate classification of StockTwits messages. Therefore, to determine the consistency of the manual classification of two independent coders, inter-rater agreement using Kappa statistics is used to perform agreement analysis that controls for the agreement expected based on chance alone. Table A presents the inter-rater agreement of the two coders based on Kappa statistics. As can be seen from Table A, the Kappa value for StockTwits data is 0.814 (81.4%) and the P value is 0; this is very small, indicating that the Kappa value is statistically significant. A Cohen's Kappa of more than 80% represents a very strong agreement and confirms higher inter-rater reliability (0.81) (Landis and Koch, 1977)

Table A: Inter-rater Agreement Test Using Kappa Statistics				
	Value	Asymp. Std. Error^a	Approx. T^b	Approx. Sig.
Measure of Agreement Kappa	0.814	0.009	60.264	0.000
N of Valid Cases	2892			

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.

Appendices

Appendix II

Research Licence Agreement

THIS RESEARCH AGREEMENT (this "Agreement") is entered into and effective as of the later of the dates set forth on the signature page of this Agreement (the "Effective Date") between Stocktwits, Inc., with its principal place of business at 1307 Ynez Place, Coronado, CA ("Company") and Brunel University, with its principal place of business at Kingston Lane, Uxbridge, Middlesex UB8 3PH, UK ("Licensor").

Subject to the terms and conditions of this Agreement and the StockTwits Terms of Use, located at www.stocktwits.com/terms Company grants Licensor an exclusive, non-transferable, non-sub licensable, revocable (as expressly provided herein) licensed during the term of the Agreement to access and use StockTwits, Inc.'s short-form real-time end user updates, end user profile information, and other content (collectively, the "Content" or "Company Materials"). Licensor acknowledges and agrees that the Company Materials shall be used, and are solely licensed, for the following purposes only:

1.1 Analyzing the Company Materials for academic financial research, provided that the research does not publicly display the Company Material themselves; Use of Company Materials for research resulting in possible publication of research findings, provided that: Research based upon or making reference to Company Materials used in research clearly attributes to Company the creation, ownership and provision of data to research effort and thanks Company for assistance and support.

In no way does Licensor charge or otherwise receive compensation specifically for access to Company Materials;

Licensor does not develop any commercial or paid products or services containing research based on Company Materials; and

(d) In publishing the research Licensor provides solely visual, not programmatic access to Company Materials. Licensor may not allow third parties to export Company Content to a data store of any kind as part of accessing the research.

Restrictions. Except as otherwise expressly approved in this Agreement or by Company in writing, Licensor may not: (a) syndicate, sell, sublicense, lease, rent, loan, lend, transmit (network or otherwise) distribute or transfer the Company Materials in any manner, including without limitation through an application programming interface ("API"), to any third party; (b) create derivative works of the Company Content with the exception of research based upon Company Materials; (c) create a service intended to replace the Company's service; (d) remove, obscure or alter any copyright notices, trademark or other proprietary rights notices affixed to or contained within the Company Materials; (e) interfere with, modify, disrupt or disable features or functionality, including without limitation any such mechanism used to restrict or control the functionality of the Service or the Company's service. For the purposes of this Section, "syndicate" includes, but is not limited to (i) delivering the Company Material to a client or third party via an API, or (ii) making the Company Materials available for download by a client or any third party, whether within Licensor's products or services or otherwise.

3. Terms of Service. Licensor shall comply with Company's terms of use located at <http://stocktwits.com/terms>

4.

Term and Termination. This Agreement will commence on the Effective Date and will expire on the one (1) year anniversary of the Effective Date. Either party may terminate this Agreement immediately if the other party breaches this Agreement and does not cure such breach within fifteen (15) days.

This the Agreement constitutes the entire understanding and agreement of the parties hereto and supersedes all prior written or oral and all contemporaneous oral agreements, understandings and negotiations among the parties hereto.


Appendices

IN WITNESS WHEREOF, the parties hereto have executed this Addendum as of the Addendum Effective Date.

STOCKTWITS, INC.

LICENSOR

By: _____
Name: _____ Name: Adam Bell

By:  _____

Title: _____ Title: Head of Contracts & IP

Date: _____ Date: 19th February 2013

Appendices

Appendix III

StockTwits Application Programming Interface (API) Schema

This schema describes the full StockTwits partner level

firehoseendpoint:<http://stocktwits.com/developers/docs/api#streams-all-docs>

Message Data

id	Stocktwits unique identifier for the message
body	Message content
created_at	Date when the message was created
user: id	Stocktwits unique identifier for the user. Messages only have one user
user: username	Username of the user
user: name	Full name of the user
user: avatar_url	Path to the users avatar
user:avatar_url_ssl	SSL path to the users avatar
user: identity	The type of user, either "Official", "User"

Appendices

user:classification	The users classification, if identity is “Official” the classification is either “ir” for the companies Investor Relation department or “pro” for Professionals and Analysts that are designated on StockTwits. An identified “user” can be classified as “suggested” as a consistent respected contributor to StockTwits
user: followers	The number of people that are following this user
user: following	The number of people this user is following
user: ideas	The number of shared ideas
user:following_stocks	The number of stocks the user is following
user: bio	The users self described biography
user: website_url	A link provided by the user
user:trading_strategy: assets_frequently_traded	The users self described trading strategy describing the users assets frequently traded. Can be any: “Equities”, “Options”, “Forex”, “Futures”, “Bonds”, “PrivateCompanies”
user:trading_strategy:	The users self described trading strategy

Appendices

approach	describing the users approach. Can be one of the following: "None", "Technical", "Fundamental", "Global Macro", "Momentum", "Growth", "Value"
user:trading_strategy:holding_period	The users self described trading strategy describing the users holding period. Can be one of the following: "None", "Day Trader", "Swing Trader", "Position Trader", "Long Term Investor"
user:trading_strategy:experience	The users self described trading strategy describing the users experience. Can be one of the following: "None", "Novice", "Intermediate", "Professional"
source: id	Message source unique identifier. Source is which application the message has originated from. Messages only have one source
source: title	The title of the source application
source: url	Link to the source application
symbols: id	StockTwits symbol internal unique identifier. Messages can have more than one symbol. Daily list of symbols can be downloaded here by date: http://stocktwits.com/symbol-sync/2013-01-30.c sv

Appendices

symbols: symbol	Public ticker symbol
symbols: title	Full public title of the tickersymbol
symbols:exchange	Stock exchange the ticker symbol resides on
symbols: sector	Sector for the ticker symbol. Sector list can be downloaded here: http://stocktwits.com/sectors/StockTwits-sectors-industries.csv
symbols:industry	Industry for the ticker symbol. Industry list canbedownloaded here: http://stocktwits.com/sectors/StockTwits-sectors-industries.csv
symbols:trending	True or false flag if the ticker symbolwastrending at the time of the messagecreation
entities:sentiment	Entities are optional. User specified sentiment at time of message creation. If sentiment is set within the message this will be either 0 - "bullish" or 1- "bearish"
entities: chart:thumb	Entities are optional. Path to the charts thumbnail image
entities: chart:original	Entities are optional. Path to the charts original image

Appendices

entities: chart: url	Entities are optional. URL to the chart page on StockTwits
conversation:parent_message_id	Conversation are optional. If there is a conversation The parent message for the StockTwits unique identifier
conversation:in_reply_to_message_id	Conversation are optional. If the message is a reply to another message the StockTwits unique identifier is represented
conversation:parent	Conversation are optional. True or false value if the message is the parent message that started the conversation
conversation:replies	Conversation are optional. Number of replies at the time of the message creation

Appendices

Appendix IV

Diagnostic Tests

The preliminary step for specifying the VAR model is to examine the economic variables for different sets of diagnostic testing procedures to validate the use of the VAR model. Those diagnostic tests are as follows: co-integration, stationarity, autocorrelations, normality and heteroscedasticity.

a) Cointegration Tests

Co-integration is a statistical technique used in economic time series that has been of central concern in the finance and economics literature. It is a concept that has evolved since the 1980s (Engle and Granger, 1987) and was successfully developed by Johansen (1988 and 1991) to examine the long-run predictability and association between the series in an equation. However, economics researchers such as Park and Phillips (1989) and Sims et al. (1990) demonstrated the invalidity of hypothesis testing in the level of VARs if the variables are said to be co-integrated. One can estimate a VAR model if, and only if, the variables are not co-integrated (no long-run association among them); otherwise, one would specify a Vector Error Correction Model (VECM). Therefore, testing for co-integration among the time series is essential before estimating the VAR model. Several tests for co-integration in time series are available. For example, the Johansen likelihood ratio test of co-integration (developed by Johansen, 1988, 1991) specifies an unrestricted Vector Autoregressive (VAR) model of order k with $(n \times 1)$ endogenous variables integrated of the same order (i.e., $I(1)$) forced by a vector of $(n \times 1)$ of independent Gaussian errors. Another test was called the residual-based co-integration test developed by Gregory and Hansen (1996). Unlike the Augmented Dickey-Fuller (ADF) test, which assumes constant parameters where the power of these tests is known to be very low in the presence of any structural changes (Campos et al., 1996; Gregory and Hansen, 1996), the Hansen test proved powerful by allowing for a structural change in the relationship between variables. Regardless of the test being used, the co-integration test is always conducted under the null hypothesis H_0 : There is no co-integration when the null hypothesis is rejected based on the decision criterion that the P value is significant at the 95% confidence interval.

Appendices

b) Stationarity Test

Broadly speaking, stationarity is a process whose statistical properties do not change over time and do not follow any trends (Nason and Von Sachs, 1999). More formally, stationarity is a stochastic process whose joint probability distribution of the variables does not change when shifted over time. Several statistical techniques are available to test for the stationarity of the data in a time series. The unit root test is the most popular test of stationarity and can be employed using three methods: the Augmented Dickey–Fuller (ADF) test (Dickey and Fuller 1979), the Phillips–Perron (PP) test (Phillips and Perron 1988), and the modified Dickey–Fuller generalised least squares (DF-GLS) test (Elliott et al., 1996). To satisfy the VAR model, the variables in the model have to be stationary, such that the joint probability distribution does not change over time and does not follow any trend. In the event of non-stationarity in the data, the test has to be carried out by taking the first difference where the data are transformed to become stationary.

c) Normality

In statistics, normality is an essential assumption in measuring the variation of the data to determine whether a data set is well modelled. It determines whether the residuals under consideration are normally distributed. The absence of the normality assumption will have an impact on the statistical test of significance and may result in misspecification of the regression model, particularly in small samples (Cohen et al., 2000). To overcome non-normality of the residuals, one should check for structural changes in the data (e.g. include dummy variables) and satisfy the normality assumption to ensure that the residuals are normally distributed.

d) Heteroscedasticity Test

Heteroscedasticity is a matter of concern in the application of regression analysis that measures the variation of the variables in the regression. The existence of heteroscedasticity can violate the statistical significance test that assumes that the modelled errors are uncorrelated (normally distributed) and that the variation of the variables is constant (Field, 2009). In statistical applications of the linear regression model, one of the most important assumptions is that there is no heteroscedasticity. Failure to meet this assumption implies that the linear regression

Appendices

estimators are not Best Linear Unbiased Estimators (BLUE). The heteroscedasticity test used to test for heteroscedasticity of the residuals is always conducted under the null hypothesis H_0 : There are no cross-terms as one could reject the null hypothesis and conclude that there is no heteroscedasticity and that the residual is said to be homoscedastic, implying that that the dependent variable(s) exhibits equal levels of variance across the range of the predictor variable(s) (Hair et al., 2006).

e) Residual Autocorrelation

Autocorrelation is known as a serial correlation, which in time series models implies finding a similar pattern between series as a function of the time lag between them. The presence of the residuals autocorrelation is a frequent problem likely to be encountered by econometricians in time series data. It is common practice in the application of VAR models to ensure that the autocorrelations among residuals are not implied in the regression of the VAR model. In statistics, residuals autocorrelation can be assessed using the Durbin-Watson test (Durbin and Watson, 1950), which can be run under the null hypothesis H_0 : there are no residual autocorrelations.

Appendices

Appendix V

The performance analysis of Naive Bayes (NB), Decision Tree (RandF) and Support Vector Machine (SMO)

1- Naive Bayes Classifier

- **Accuracy Rate**

The Naive Bayes classifier was tested using both approaches: testing on full training data and testing using cross-validation with 10-folds. As can be seen from the output results shown in Table B, testing on the training data resulted in 65.53% accuracy whereas testing using cross-validation gave an accuracy of 62.76%.

Table B: Weka Summary Results for Naive Bayes Classifier (Full Training Set and 10-Folds Cross-Validation)			
Testing Method	Accuracy Rate	Correct Classified Instance	Incorrect Classified Instance
Full Training Set	65.53%	1,895	997
10-Folds Cross-Validation	62.76%	1,815	1,077

Testing directly on the training data classified 1,895 instances correctly out of 2,892, which is an accuracy of 65.53%. A very high level of accuracy is always required when testing on the full training sets because it indicates the magnitude at which the model has learnt the training data. Using the Naive Bayes classifier, the overall classification accuracy testing on the full training data was 65.53%. This is considered a good percentage giving a random chance of 33% of the three classes (buy, sell and hold) (equal probability of each class is $1/3 = 0.33 = 33\%$). Therefore, an accuracy of 65.53% is well within the error range and reveals that the Naive Bayes has learnt the training data quite accurately. Testing by 10-folds cross-validation correctly classified 1,815 instances out of 2,892, achieving an average accuracy of 62.76%. An accuracy of 62.76% is well within the desired range; as stated above, any accuracy of more than 33% is considered good.

- **Classification Accuracy by Class**

Using stratified cross-validation with 10-folds, the 'buy' class has a precision of 68.30%, a recall of 71.5% and an F-measure of 69.90%, indicating a reasonably good performance by the Naive

Appendices

Bayes model in predicting that class. The ‘hold’ class has a precision of 53.10% and a recall of 61.70%, while the F-measure shows a relative fall to 57.10%. Meanwhile, the ‘sell’ class has a precision of 61.10% but the recall is relatively low at 50.80%, causing the F-measure to drop to 55.50%. The weighted averages of the three classes are shown in the last row of Table C. The weighted averages achieve very similar results of 62.90%, 62.80% and 62.60% for precision, recall and F-measures respectively.

Table C: Classification Accuracy By Class Using Naive Bayes Classifier						
Full Training Set						
Class	True Positive	False Positives	Precision	Recall	F- Measure	ROC Area
Buy	73.30%	27.20%	70.60%	73.30%	71.90%	79.50%
Hold	66.80%	13.20%	56.50%	66.80%	61.20%	85.60%
Sell	53.50%	14.20%	64.40%	53.50%	58.40%	77.40%
Weighted Average	65.50%	20.10%	65.70%	65.50%	65.30%	80.10%
10-fold cross-validation						
Class	True Positive	False Positives	Precision	Recall	F-Measure	ROC Area
Buy	71.50%	29.50%	68.30%	71.50%	69.90%	77.10%
Hold	61.70%	14.00%	53.10%	61.70%	57.10%	83.00%
Sell	50.80%	15.60%	61.10%	50.80%	55.50%	75.00%
Weighted Average	62.80%	21.80%	62.90%	62.80%	62.60%	77.60%

Note: True positive represents the messages correctly classified as a given class. False positives are messages classified incorrectly as a given class. Precision is the proportion of the messages truly classified in a class divided by the total messages classified as that class. Recall (also known as sensitivity) of a particular class represents the share of all messages that were classified correctly. Note that the Recall measure is equivalent to True Positive rate. F-measure is a combined measure for precision and recall and is calculated as $F = (2 * Precision * Recall) / (Precision + Recall)$. ROC Area measure is one of the most important measures of Weka output. It exemplifies the performance of the classification model through the trade-off between the classifier sensitivity (TP_{rate}) and specificity (false alarm rate FP_{rate}) where the sensitivity can only be increased with a little loss in specificity and vice versa.

With 10-fold cross-validation, the Naive Bayes classifier shows that the area under the curve (AUC) filled up 77.10%, 83.00% and 75.00% for the buy, hold and sell classes respectively. The area under the curve is a very reasonable measure that can provide the classification accuracy. The closer the area under the curve of a given class is to 100%, the better the classifier will be in predicting that particular class. From the ROC Area results shown in Table C, it can be seen that the AUC for the hold class reported the highest measure of 83.00% compared to the other two classes in the dataset, indicating that the Naive Bayes classifier is very good at predicting

Appendices

instances in the hold class.

- **Confusion Matrix**

Table D shows the confusion matrix for the Naive Bayes Classifier as it appears at the bottom of the output results for both methodologies when testing using the full training set and testing using the 10-folds cross-validation test. As can be seen from Table D, when using the full training test, the Naive Bayes model shows a prediction accuracy of 73.30%, 66.80% and 53.50% for buy, hold and sell classes respectively. While showing good classification accuracy for the buy and hold classes, it is harder for the model to predict the messages of the sell class as they are confused with the buy and hold messages. Using the 10-folds cross-validation, the results clearly show that a total of 973 out of 1,361 instances were correctly classified as buy, which is an accuracy of 71.50% (indicated by the true positive rate). Meanwhile, 364 out of 590 instances were correctly classified as hold messages, an accuracy of 61.70%, and only 478 out of 941 instances were classified accurately in the sell class, which gives an accuracy of 50.80%. Obviously, the buy class is the most successfully predicted by the model in both methodologies compared with the other two classes.

Table D: Classification Accuracy (Confusion Matrix) for Naive Bayes Classifier			
Full Training Sets			
Classified As	Buy	Hold	Sell
Buy	998	160	203
Hold	12	394	75
Sell	295	143	503
10 -Folds Cross-Validation			
Classified As	Buy	Hold	Sell
Buy	973	168	220
Hold	142	364	84
Sell	309	154	478

Note: Each element in the matrix is a count of instances. The rows represent the true classified instances (messages) of a given class while the columns represent the predicted instances of that class.

2- Decision Tree (RandF) Classifier

- **Accuracy Rate**

The Random Forest decision tree classifier was tested using two methodologies: testing on full training data and testing using cross-validation with 10-folds. From the output results shown in

Appendices

Table E, testing on the training data produced 97.72% accuracy whereas testing using cross-validation gave an accuracy of 66.70%. The complete results along with a discussion are as follows:

Testing Method	Accuracy Rate	Correctly Classified Instances	Incorrectly Classified Instances
Full Training Set	97.72%	2,826	66
10-Folds Cross-Validation	66.70%	1,929	963

Testing directly on the training data classified 2,826 instances correctly out of 2,892, yielding an accuracy of 97.72% using the Random Forest DT classifier. Such a high accuracy of 97.72% achieved by the model signifies that the built Random Forest model has learnt the training data extremely well. It is considered an excellent accuracy level. Since there were three classes in the target variables (buy, sell and hold), any percentage greater than 33% (meaning the probability of occurrence of each class is $1/3 = 0.33 = 33\%$) has to be considered good. With the Random Forest, the 10-folds cross-validation experiments achieved an accuracy of 66.70%, where 1,929 instances were correctly classified out of 2,892. An accuracy of 62.76% is well within the desired range, giving a random chance probability of 33% for each class.

- **Classification Accuracy by Class**

Table F shows the classification accuracy by class for the Random Forest Tree classifier model where the cost-insensitive measures (such as precisions, recall and F-measures) as well as the cost-sensitive measures (e.g. ROC Area) for each class are reported. Using stratified cross-validation with 10-folds, the buy class has a precision of 69.60%, a recall of 78.40% and an F-measure of 73.70% indicating a very good performance by the Random Forest model in predicting that class. The hold class has a precision of 61.60%, a recall of 63.90% and an F-measure of 62.70%. The sell class has a precision of 65.00% but the recall is relatively low at 51.50%, causing the F-measure to drop to 57.50%. The weighted averages of the three classes are shown in the last row of the Table, indicating very similar results of 65.50%, 66.70% and 66.20% for precision, recall and F-measures respectively.

Appendices

Table F: Classification Accuracy By Class Using Decision Tree Classifier (Random Forest)						
Full Training Set						
Class	True Positive	False Positives	Precision	Recall	F-Measure	ROC Area
Buy	99.50%	2.90%	96.80%	99.50%	98.10%	99.90%
Hold	97.10%	0.40%	98.30%	97.10%	97.70%	99.90%
Sell	95.50%	0.60%	98.80%	95.50%	97.10%	99.80%
Weighted Average	97.70%	1.70%	97.70%	97.70%	97.70%	99.90%
10-Folds Cross-Validation						
Class	True Positive	False Positives	Precision	Recall	F-Measure	ROC Area
Buy	78.40%	30.50%	69.60%	78.40%	73.70%	80.00%
Hold	63.90%	10.20%	61.60%	63.90%	62.70%	85.90%
Sell	51.50%	13.40%	65.00%	51.50%	57.50%	75.70%
Weighted Average	66.70%	20.80%	66.50%	66.70%	66.20%	79.80%

See notes to Table B.

With 10-folds cross-validation, the Random Forest tree shows that the area under the curve (AUC) filled up at 80.00%, 85.90% and 75.70% for the buy, hold and sell classes respectively. The area under the curve is a very reasonable measure that can indicate the classification accuracy. The closer the area under the curve of a given class is to 100%, the better the classifier will be in predicting that particular class. Although the Random Forest tree shows relatively good classification accuracy indicated by the AUC, which scored > 75.00% in all three classes, the AUC for the hold class reported the highest measure of 85.00% compared to the buy and sell classes. This means that the model accuracy indicated by the AUC for the hold class outperforms the accuracy of that model reported for the other two classes using the same evaluation measure (AUC).

- **Confusion Matrix**

The confusion matrix is one of the most important performance evolution measures of machine learning models. As can be seen from Table G, using the full training test the Random Forest Tree model shows prediction accuracies of 99.50% (1,354/1361), 97.10% (573/590) and 95.50% (899/941) for buy, hold and sell classes respectively. Note that the number of manually classified tweets are 1,361, 590 and 941 for the buy, hold and sell classes respectively. The Random Forest

Appendices

Tree model shows excellent classification accuracy for all three classes when testing on the full training data.

Table G: Classification Accuracy (Confusion Matrix) for Decision Tree (Random Forest) Classifier			
Full Training Sets			
Classified As	Buy	Hold	Sell
Buy	1,354	2	5
Hold	11	573	6
Sell	34	8	899
10-Folds Cross-Validation			
Classified As	Buy	Hold	Sell
Buy	1,067	117	177
Hold	129	377	84
Sell	338	118	485

Note: See notes to Table B

The second part of Table G presents the confusion matrix for the 10-folds cross-validation, showing what classification the instances (messages) from each class received when used as testing data. For example, the buy class has 1,361 buy instances ($1,361 = 1,067 + 117 + 177$) in our training example, of which the classification model correctly classified 1,067 as buy but incorrectly identified 294 instances ($294 = 117 + 177$) as hold or sell. Therefore, 1,067 instances are the true positives and 294 instances are the false negatives of the buy class. Meanwhile, for the hold class, we have 590 hold instances in the training sample, of which the model produced 377 correctly classified instances, while 213 instances ($213 = 129 + 84$) were misclassified of which 129 and 84 instances were incorrectly assigned to the buy and sell classes respectively. Hence, 377 and 213 instances were the true positives and the false negatives respectively of the hold class. On the other hand, for the sell class, the model correctly classified 485 instances (indicating the true positive instances) while 338 and 118 instances were misclassified (indicating the false negative instances) and were assigned to the buy and hold classes respectively. Overall, the Random Forest decision tree model shows the highest classification accuracy of the messages belonging to the buy class but was less successful at predicting the messages in the sell class.

Appendices

3. Support Vector Machines (SMO)

- **Accuracy Rate**

The Sequential Minimal Optimisation (SMO) algorithm of SVMs was tested using two methodologies: testing on full training data and testing using cross-validation with 10-folds. From the output results shown in Table H, testing on the training data produced 68.78% accuracy whereas testing using cross-validation gave an accuracy of 65.25%. The complete results along with a discussion are as follows:

Testing Method	Accuracy Rate	Correct Classified Instance	Incorrect Classified Instance
Full Training Set	68.78%	1,889	903
10-Folds Cross-Validation	65.25%	1,887	1,055

Testing on the full training sets correctly classified 1,989 instances out of 2,892, achieving an accuracy of 68.78%. The higher the classification accuracy of the model, the better the model has learnt the trading data with high accuracy. An accuracy of 68.78% is considered well within the desired range. Testing with 10-folds cross-validation achieves an average of 1,887 correct classification instances out of 2,892, yielding an accuracy of 65.25%. An accuracy of 65.25% is considered well within the desired range.

- **Classification Accuracy by Class**

Table I provides the classification accuracy by class for the SMO classifier model where the cost-insensitive measures (such as precisions, recall and F-measures) as well as the cost-sensitive measures (e.g. ROC Area) are reported independently for each class. Using stratified cross-validation with 10-folds shown in the second half of the Table I, the buy class reported a precision of 64.00% and a recall of 88.80%, indicating a reasonably good performance by the SMO model in predicting the buy class. The buy class reported an F-measure of 72.90%. The hold class has a precision of 63.80%, but a low recall of 48.60%, causing the F-measure to

Appendices

decline to 55.20%. The sell class reported a high precision of 69.90% but the recall is much lower at 47.40%, causing the F-measure to yield a small percentage of 56.50%. The weighted averages of the three classes are shown in the last row of the Table. The weighted averages achieved by the model are 65.90%, 65.20% and 64.00% for precision, recall and F-measures respectively. It is apparent from the Table below that the AUC for the buy class filled in 72.60%. For the hold class, the AUC occupied 80.10%, which indicates a very good accuracy of SMO in predicting instances of the hold class. For the sell class the AUC filled about 74.10% of the area under the ROC curve. As with the Naive Bayes and Random Forest, the SMO model measured by the ROC curve has recorded very good accuracy in predicting the hold instances, and it is considered reasonably good at predicting instances of buy and sell classes.

Table I: Classification Accuracy By Class Using Support Vector Machines (Sequential Minimal Optimisation)						
Full Training Set						
Class	True Positive	False Positives	Precision	Recall	F- Measure	ROC Area
Buy	87.40%	39.30%	66.40%	87.40%	75.50%	75.80%
Hold	54.60%	5.90%	70.30%	54.60%	61.50%	83.40%
Sell	50.80%	8.50%	74.20%	50.80%	60.30%	74.10%
Weighted Average	68.80%	22.40%	69.80%	68.80%	67.70%	76.80%
10-Folds Cross-Validation						
Class	True Positive	False Positives	Precision	Recall	F-Measure	ROC Area
Buy	84.80%	42.50%	64.00%	84.80%	72.90%	72.60%
Hold	48.60%	7.10%	63.80%	48.60%	55.20%	80.10%
Sell	47.40%	9.80%	69.90%	47.40%	56.50%	71.10%
Weighted Average	65.20%	24.60%	65.90%	65.20%	64.00%	73.60%

Note: Note: See notes to Table B

- **Confusion Matrix**

Table J shows the confusion matrix for Support Vector Machines (SMO). As can be seen from the Table, using the full training test the SMO model shows prediction accuracies of 87.40%, 54.60% and 50.80% for the buy, hold and sell classes respectively. While the SMO model shows very good classification accuracy for the buy class, it is less successful at predicting the messages of the hold and sell classes. Generally, the total number of correct classifications is lower in all three classes. A total of 1,154 out of 1,361 instances were correctly classified as buy, which is an accuracy of 84.80%. However, only 287 out of 590 instances were correctly

Appendices

classified as hold messages, an accuracy of 48.60%, and only 446 out of 941 instances were classified accurately in the sell class, an accuracy of 47.40%. Obviously, the buy class is the class predicted most successfully by the SOM model in both methodologies while the hold and sell classes are harder to predict.

Table J: Classification Accuracy (Confusion Matrix) for Support Vector Machines (Sequential Minimal Optimization)			
Full Training Sets			
Classified As	Buy	Hold	Sell
Buy	1189	71	101
Hold	203	322	65
Sell	398	65	478
10-Folds Cross-Validation			
Classified As	Buy	Hold	Sell
Buy	1154	91	116
Hold	227	287	76
Sell	423	72	446

Note: See notes to Table B

Appendices

Appendix VI

An Extracted Screen of Visualized Decision Tree

These two figures show an exemplary screen of a visualised decision tree (DT) model of StockTwits data. The decision tree is structured such that each node in the tree is connected through leaves to another decision node and both are connected to a leaf node holding the class prediction. The prediction class may take three possible states $c = \{\text{Sell, Buy, Hold}\}$. An inductive (If-Then) rule is created for each path from the root to the leaf by which the trading decision is predicted. The visualised tree displays all possible trading decisions represented either by an individual term or pair-wise combinations of terms. Closer inspection of Figure A reveals that a number of trading decision guidelines can be extracted from the DT model, based on the (if-then) rule.

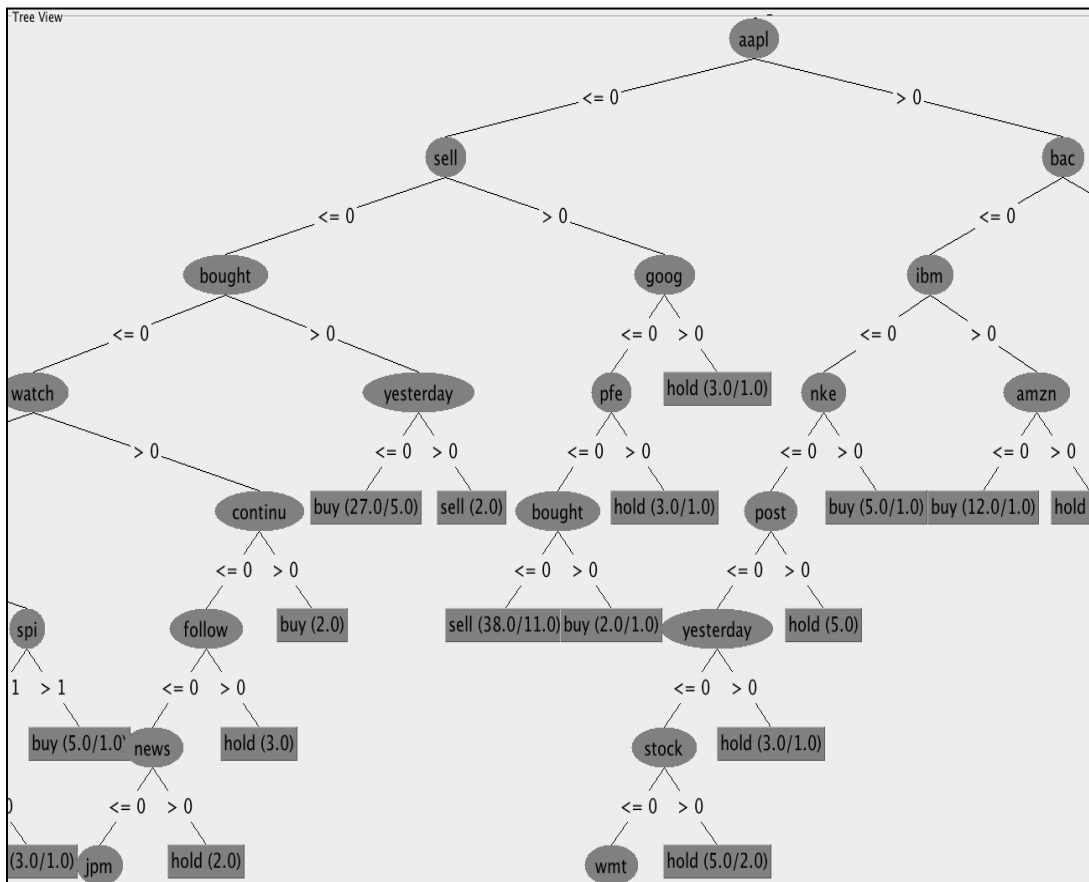


Figure A: a screen shot of a visualised DT model: focus on the decision nodes: “apl” and “sell”

From the extracted visualised tree model graphed in Figure A, we can see that the decision class (hold) is connected, through leaves, to the words (apl, bac, ibm, nke, post, yesterday and stock).

Appendices

This indicates that these words are the most relevant words that best classify “hold” messages. Each term is indicated by a node in the tree and connected through leaves to one of the three decision classes (sell, buy or hold). For example, *if* the combined terms such as “apl+ ibm+ amzn”, “apl+ post” “apl+ yesterday” and “apl+ stocks” appear in a tweet message, *then* the decision recommended by that investor is to hold that particular stock. The decision nodes of an attribute might be affected differently depending on whether that attribute appears individually or in combinations with other attributes in the tree. For example, one might view another path in the tree such as the decision node “sell” where it is connected to other decision nodes in the tree such as “goog, pfe and bought”. If the term “sell” appears alone in a tweet message, it will strongly signify a sell decision indicated by the predicted decision class {sell} at the end of the leaf node that holds a sell class. However, the decision node “sell” indicates a precisely inverse decision when it is connected to the decision node “bought”; i.e. if the combined terms “sell+ bought” appear in a tweet, this will signal a buying decision indicated by the decision class “buy” at the end of the tree root. Meanwhile, another decision might be recommended when the term “sell” is combined with the term “pfe”, showing a holding position.

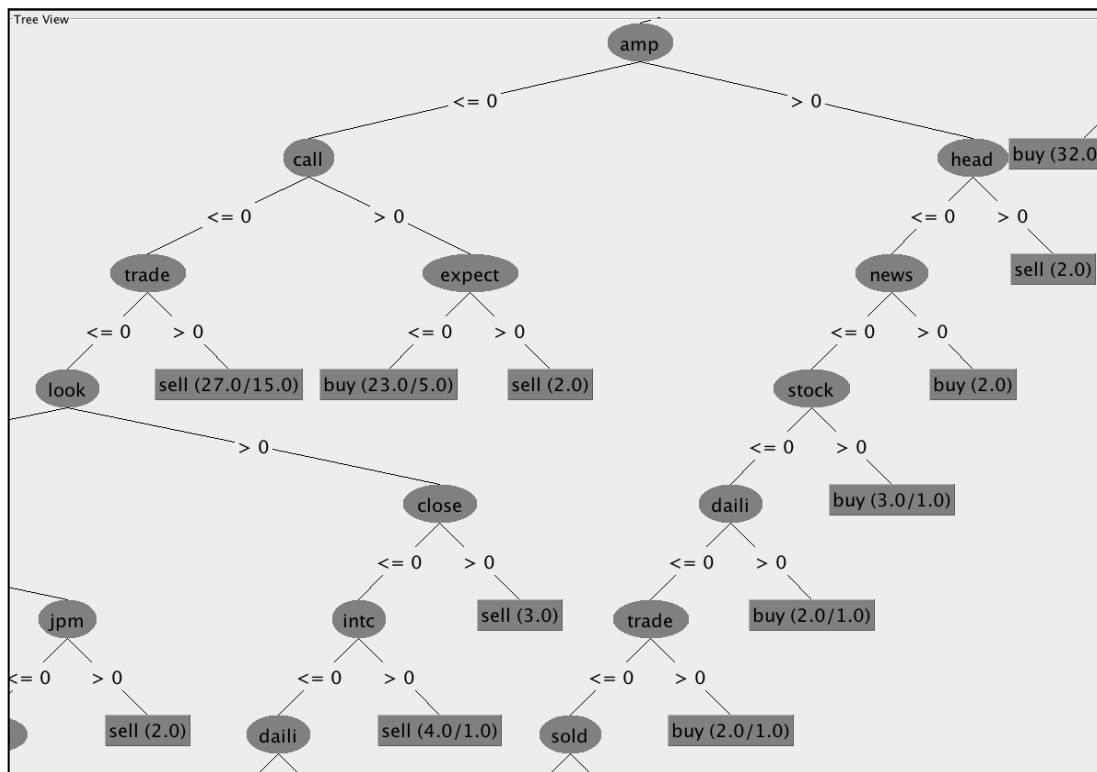


Figure B: A screenshot of a visualised DT model: focus on the decision nodes: “amp” “look” and “call”

Appendices

Exploring different interactions of the combined appearance of the terms in tweet postings, another screenshot is presented in Figure B. The decision node “amp” is connected through leaves to different decision nodes (such as head, news, stock, daily, trade and sold). Interestingly, all the pair-wise combinations of attributes connected with the term “amp” show buying decisions except for the combined appearance of “amp+ head” where the predicted class indicates a sell decision. The decision class “call” is connected to another decision node “expect”, and their combined appearance signifies a sell decision despite the dominant buying decisions signalled by the individual appearance of the term “call” in a tweet message. Another branch of tree might be explored by viewing the term “look”, whose combined appearance with the other terms in the tree (i.e. close and intc) indicate a sell signal to stock market participants.

APPENDICES REFERENCES

- Antweiler, W. and Frank, M.Z. 2004b. Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance*, **59**(3), pp. 1259-1294.
- Artstein, R. and Poesio, M. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, **34**(4), pp. 555-596.
- Burns Alvin, C. and Bush Ronald, F. 1995. Marketing Research. *Eaglewood Cliffs: Prentice Hall, New Jersey*, pp. 3-5.
- Campos, J., Ericsson, N.R. and Hendry, D.F. 1996. Cointegration tests in the presence of structural breaks. *Journal of Econometrics*, **70**(1), pp.187-220.
- Churchill, G.A. and Iacobucci, D. 2009. *Marketing research: methodological foundations*. Cengage Learning.
- Cohen, L., Manion, L. and Morrison, K. 2000. Research Methods in Education [5 th edn] London: Routledge Falmer. *Teaching in Higher Education*, **41**.
- Craggs, R. and Wood, M.M. 2005. Evaluating discourse and dialogue coding schemes. *Computational Linguistics*, **31**(3), pp. 289-296.
- David, M. and Sutton, C.D. 2004. *Social research: The basics*. London:Sage Publication.
- Dickey, D.A. and Fuller, W.A. 1979. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association*, **74**(366a), pp. 427-431.
- Durbin, J. and Watson, G.S., 1950. Testing for serial correlation in least squares regression: I. *Biometrika*, pp. 409-428.
- Engle, R.F. and Granger, C.W. 1987. Co-integration and error correction: representation, estimation, and testing. *Econometrica: journal of the Econometric Society*, pp.251-276.
- Field, A., 2009. *Discovering statistics using SPSS*. London: Sage publications.
- Gregory, A.W. and Hansen, B.E. 1996. Residual-based tests for cointegration in models with regime shifts. *Journal of econometrics*, **70**(1), pp.99-126.
- Hair, J.F., Tatham, R.L., Anderson, R.E. and Black, W., 2006. *Multivariate data analysis*. Pearson Prentice Hall Upper Saddle River, NJ.
- Johansen, S. 1991. Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica: Journal of the Econometric Society*, pp. 1551-1580.

Appendices

Johansen, S. 1988. Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*, **12**(2), pp. 231-254.

Landis, J.R. and Koch, G.G. 1977. The measurement of observer agreement for categorical data, *biometrics*, pp.159-174

Nason, G.P. and Von Sachs, R. 1999. Wavelets in time-series analysis. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, **357**(1760), pp. 2511-2526.

Park, J.Y. and Phillips, P.C. 1989. Statistical inference in regressions with integrated processes: Part 2. *Econometric Theory*, **5**(01), pp.95-131.

Phillips, P.C. and Perron, P. 1988. Testing for a unit root in time series regression. *Biometrika*, **75**(2), pp. 335-346.

Sims, C.A., Stock, J.H. and Watson, M.W. 1990. Inference in linear time series models with some unit roots. *Econometrica: Journal of the Econometric Society*, pp.113-144.

Sprenger, T.O., Tumasjan, A., Sandner, P.G. and Welpe, I.M. 2014. Tweets and trades: The information content of stock microblogs. *European Financial Management*, **20**(5), pp. 926-957.