Essays on Financial Applications of Nonlinear Models

Wanbin Wang

A thesis submitted for the degree of Doctor of Philosophy of The Australian National University

November 2017

Declaration

I, declare that this thesis, is my original work and that it contains no material previously published or written by another person, except where due acknowledgement is made in the text.

_____

Wanbin Wang

Acknowledgements

I am greatly indebted to my primary supervisor, Dr Kin–Yip Ho, who has continually lent me his guidance and support throughout my PhD studies, inspiring me to strive for a high level of excellence in my research, supporting me to attend conferences and helping me to develop academic networks. He offered his detailed insights into every draft of this thesis. This thesis would not have been completed without his inspiration, encouragement and support.

I am extremely grateful to my supervisor, Dr Wai–man (Raymond) Liu. His inspiring guidance and advice have tremendously improved this thesis, as well as my research abilities. I also wish to thank my another supervisor, Dr Zhaoyong Zhang, for his unfailing support, valuable suggestions and comments on my work.

I gratefully acknowledge the financial support I have received from the Australian National University College of Business and Economics and the Research School of Finance, Actuarial Studies and Statistics (RSFAS). This not only allowed me to pursue my doctoral studies, it also placed me within a network of other committed scholars. I extend my gratitude to other members of the RSFAS for their constant support.

My achievements would have been impossible without my wife's patience, tolerance and unending support and love. My son, Yipu Wang, and my daughters, Kelly Wang and Ellie Wang, have been the joys of my life. I also want to thank my mother, sisters and brother for their unfailing love, encouragement and support.

Abstract

In this thesis, we examine the relationship between news and the stock market. Further, we explore methods and build new nonlinear models for forecasting stock price movement and portfolio optimization based on past stock prices and on one type of big data, news items, which are obtained through the RavenPack News Analytics Global Equities editions.

The thesis consists of three essays. In Essay 1, we investigate the relationship between news items and stock prices using the artificial neural network (ANN) model. First, we use Granger causality to ascertain how news items affect stock prices. The results show that news volume is not the Granger cause of stock price change; rather, news sentiment is. Second, we test the semi–strong form efficient market hypothesis, whereas most existing research testing efficient market hypothesis focuses on the weak–form version. Our ANN strategies consistently outperform the passive buy–and–hold strategy and this finding is apparently at odds with the notion of the efficient market hypothesis. Finally, using news sentiment analytics from RavenPack Dow Jones News Analytics, we show positive profitability with out–of–sample prediction using the proposed ANN strategies for Google Inc. (NASDAQ: GOOG).

In Essay 2, we expand the utility of the information from news volume and news sentiments to encompass portfolio diversification. For the Dow Jones Industrial Average (DJIA) components, we assign different weights to build portfolios according to their weekly news volumes or news sentiments. Our results show that news volume contributes to portfolio variance both in–sample and out–of–sample: positive news sentiment contributes to the

portfolio return in–sample, while negative contributes to the portfolio return out–of–sample, which is a consequence of investors overreacting to the news sentiment. Further, we propose a novel approach to portfolio diversification using the k–Nearest Neighbors (kNN) algorithm based on the idea that news sentiment correlates with stock returns. Out–of–sample results indicate that such strategy dominates the benchmark DJIA index portfolio.

In Essay 3, we propose a new model called the Combined Markov and Hidden Markov Model (CMHMM), in which observation is affected by a Markov model and an HMM (Hidden Markov Model) model. The three fundamental questions of the CMHMM are discussed. Further, the application of the CMHMM, in which the news sentiment is one observation and the stock return is the other, is discussed. The empirical results of the trading strategy based on the CMHMM show the potential applications of the proposed model in finance.

This thesis contributes to the literature in a number of ways. First, it extends the literature on financial applications of nonlinear models. We explore the applications of the ANNs and kNN in the financial market. Besides, the proposed new CMHMM model adheres to the nature of the stock market and has better potential prediction ability. Second, the empirical results from this dissertation contribute to the understanding of the relationship between news and the stock market. For instance, our research found that news volume contributes to the portfolio return and that investors overreact to news sentiment—a phenomenon that has been discussed by other scholars from different angles.

Table of Contents

List of Tables

List of Figures

Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Network
APT	Arbitrage Pricing Theory
CMHMM	Combined Markov and Hidden Markov Model
CAPM	Capital Asset–Pricing Model
DJIA	Dow Jones Industrial Average, or Dow Jones 30
DSS	Daily Sentiment Score
EMH	Efficient–Market Hypothesis
EM	Expectation Maximization
ENS	Event Novelty Score
ESS	Event Sentiment Score
HMM	Hidden Markov Model
IPO	Initial Public Offering
kNN	k–Nearest Neighbour
MPT	Modern Portfolio Theory
NASDAQ	National Association of Securities Dealers Automated Quotation
NYSE	New York Stock Exchange
PNN	Probabilistic Neural Network
RPNA	RavenPack News Analytics
SIRCA	Securities Industry Research Centre of Australasia
TRNA	Thomson Reuters News Analytics
TRTH	Thomson Reuters Tick History
UTC	Coordinated Universal Time

Chapter 1:
Introduction

1.1 Research background

The stock market is one of the most important ways for companies to raise money and has become an integral part of the global economy. Investing in stocks has been one of the most popular investments for investors. However, there is always some risk to investment in the stock market as it is very hard to predict stock price movement. Forecasting stock price movement is extremely challenging as the stock market is essentially dynamic, nonlinear, complicated and nonparametric in nature.

Researchers have shown that stock price fluctuations depend on many factors, including equity (Lucas and McDonald, 1990, Brav et al., 2000), interest rates (Christie, 1982, Flannery and James, 1984, Alam and Uddin, 2009), cash flows (Sloan, 1996, Chen et al., 2013), insider information (Kyle, 1985, Wang and Wang, 2017), unexpected extreme news (Chan, 2003, Asgharian et al., 2011), prescheduled earnings announcements (Jennings and Starks, 1986, Skinner, 1994, Su, 2003), political events (Kim and Mei, 2001, Amihud and Wohl, 2004, Jensen and Schmith, 2005), and corporate takeovers (Malatesta and Thompson, 1985, Franks and Harris, 1989, Pound and Zeckhauser, 1990) etc.

A large amount of research has been published on, continues to build a prediction model for and uses different techniques to predict the stock market (Park and Irwin, 2007, Atsalakis and Valavanis, 2009, Tziralis and

Tatsiopoulos, 2012, Nazário et al., 2017) or build portfolios (Markowitz, 1952, Konno and Yamazaki, 1991, Paranjape-Voditel and Deshpande, 2013). Traditional forecasting research has employed statistical time series analysis techniques such as autoregression moving average (Rogalski, 1978, Atsalakis and Valavanis, 2009, Taylor, 2011) and regression models (Cutler et al., 1989, Schwert, 1989, Tsai, 2012). In recent years, with successful applications of Artificial intelligence (AI) techniques across a wide range of fields including medical diagnosis (Szolovits et al., 1988, Kononenko, 2001, Ramesh et al., 2004, Fieschi, 2013), robot control (Nguyen-Tuong and Peters, 2011, Ingrand and Ghallab, 2014, Siciliano and Khatib, 2016), online and telephone customer service (Hui and Jha, 2000, Zeinalizadeh et al., 2015, Rodríguez et al., 2016), remote sensing (Estes et al., 1986, Tuia et al., 2014, Lary et al., 2016) and toys (Lund, 2003), numerous stock prediction systems based on AI techniques, including artificial neural networks (ANN), fuzzy logic, the hybridization of ANN and fuzzy systems and support vector machines have been proposed (Park and Irwin, 2007, Atsalakis and Valavanis, 2009).

Atsalakis and Valavanis (2009) summarise the applications of some intelligent techniques to forecast stock market indexes and stock prices. These techniques include ANNs, genetic algorithms, autoregressive moving average models and autoregressive integrated moving average models. According to Atsalakis and Valavanis (2009), the number of input variables used in each model differs. In general, the average number of input variables is between four and ten while the most commonly used inputs are the stock (index) opening price, closing price, and highest and lowest daily values. The performance measures used by different authors may be classified as

statistical measures and non–statistical measures. Statistical measures include the root mean square error (RMSE), the mean absolute error (MAE) and the mean squared prediction error (MSPE), as well as statistical indicators such as the autocorrelation, the correlation coefficient, the mean absolute deviation, the squared correlation and the standard deviation.

In the past years, data has increased on large scales in various fields. Industries are interested in the potential of big data. The burgeoning data deluge in this era of big data heralds significant challenges for data analysis (Chen et al., 2014). Nearly all major companies, including EMC, Oracle, IBM, Microsoft, Google, Amazon and Facebook, have started their big data projects. Many national governments have likewise been highly attentive to big data. In March 2012, the Obama administration announced a USD 200 million investment to launch the 'Big Data Research and Development Plan'. The spotlight has also been on big data in academia. In 2011, Science launched a special issue (vol. 331, no. 6018) titled 'Dealing with Data' on the key technologies of data processing in big data.

Big data can improve the productivity and competitiveness of enterprises and the public sector, and create huge benefits for consumers. According to McKinsey and Company reports (Manyika et al., 2011), if big data could be creatively and effectively utilised, the potential value of the US medical industry may surpass USD 300 billion, thus reducing the requisite expenditure for the US. healthcare by over 8%. Farecast, an airline ticket forecast system that predicts trends and rising/dropping ranges in airline ticket prices, has saved nearly USD 50 per ticket per passenger, with its forecast accuracy as high as 75% (Mayer-Schönberger and Cukier, 2013).

Big data also provide sought-after opportunities for technical analysis in the domain of finance. A technical trading system consists of a set of trading rules that generate trading signals; for example, long, short, or out of the market. Most existing research on trading strategies only considers past stock price (Atsalakis and Valavanis, 2009). Big data provide useful information for building trading strategies. For instance, Bettman et al. (2010) reveal that message–board takeover rumours generate significant positive abnormal returns and trading volumes.

In this thesis, we consider one type of big data, news items. We use the dataset obtained from the RavenPack News Analytics (RPNA) Dow Jones Edition, which has been widely used by other researchers (Mitra and Mitra, 2011, Shi and Ho, 2015, Akbas et al., 2016, Shi et al., 2016b). RavenPack systematically tracks and analyses information on more than 2,200 government organisations, 138,000 key geographical locations, 150 major currencies, 80 traded commodities and over 30,000 companies. It is a comprehensive database covering more than 1,200 types of firm–specific and macroeconomic news events. Among its many benefits, RavenPack delivers sentiment analysis and reveals the event data that are most likely to affect financial markets and trading around the world. All relevant news articles about entities are classified and quantified according to their sentiment, relevance, topic, novelty and market effect. The more details introduction of RavenPack is given in the Appendix.

In general, linear models are not sufficiently adequate for describing and predicting the number of features associated with the stock market. In this thesis, we consider using nonlinear models to describe and predict the stock

market. That is, we examine the relationship between news and the stock market. Further, we explore methods for forecasting stock price movement and portfolios using nonlinear models, specifically, ANN models, k–Nearest Neighbor (kNN) algorithm and Markov models/hidden Markov models (HMMs). The purpose and approach of each of the three essays that constitute this thesis are summarised below.

1.2 Essay one

AI techniques are changing our world with successful applications in different domains. Among these AI techniques, the ANN is one of the most popular. The structure of the ANN model mimics the human brain and nervous system (Hill et al., 1994, Zhang et al., 1998, Bahrammirzaee, 2010). ANN is a data–driven modelling approach and a nonlinear nonparametric model. ANNs utilise data and let the data determine the structure and parameters of a model. In Essay one, we explore the relationship between news items and the stock return using the ANN model.

First, we seek to discover what effects stock price movement using the Granger causality test (Granger, 1969, Granger, 1988), a statistical hypothesis test for determining whether one–time series is useful in forecasting another. Our results show that news volume is not the Granger cause of stock price change; news sentiment is the Granger cause of stock price change.

Second, we wish to test the semi–strong form efficient market hypothesis, whereas most existing research on testing efficient markets hypothesis focuses on the weak–form version. According to the efficient markets hypothesis, it is impossible to ‘beat the market’ as market prices reflect all

relevant information. The existence of statistical arbitrage and profitable trading strategies are incompatible with market efficiency. We consider a broad range of news releases in the stock market, build ANN trading strategies using news items and stock return as inputs, and perform out–of–sample forecasting. The news releases are extracted from the unique RavenPack News Analytics database that monitors over 1,000 types of events ranging from different corporate actions to terrorist threats and natural disasters. We find that the ANN strategies consistently outperform the passive buy–and–hold strategy and that this finding is apparently at odds with the notion of the efficient market hypothesis.

Finally, we build a trading strategy considering a company and test the potential profitability of the ANN strategies. Using news sentiment analytics from RavenPack Dow Jones News Analytics, we develop an ANN model to predict the stock price movements of Google Inc. (NASDAQ: GOOG) and test its potential profitability using out–of–sample prediction.

1.3 Essay two

Modern Portfolio Theory (MPT) begins with the path–breaking work of Markowitz (1952). Markowitz's mean–variance optimization method suggests that it is possible to construct an 'efficient frontier' of optimal portfolios, offering the maximum possible expected return for a given level of risk.

Since Markowitz, researchers have proposed alternative portfolio theories that include additional moments such as skewness or more realistic descriptions of the distribution of returns. Others have improved Markowitz's mean–variance approach by reducing statistical errors in the sample mean

and covariance matrix.

However, the efficiency of Markowitz's mean–variance portfolio optimization method is in question. For instance, an empirical study by DeMiguel et al. (2009) evaluates the mean–variance portfolio method across seven empirical datasets and finds it leads to poor out–of–sample performances, no better than the 1/N rule in terms of Sharpe ratio, certainty–equivalent return or turnover.

The sharply increasing data deluge in the big data era presents significant challenges for portfolio diversification. In essay two, we expand the use of information from news volume and news sentiments to portfolio diversification. We discuss the possibility of the contribution of news volume and news sentiments to portfolios by assessing the performance of portfolios that are constructed according to these factors. Our results show that news volume contributes to portfolio variance both in–sample and out–of–sample; positive news sentiment contributes to portfolio return in–sample; and negative news sentiment contributes to portfolio return out–of–sample, which is a consequence of investor overreaction to news sentiment.

Further, we propose a novel approach to portfolio diversification based on the kNN algorithm. The diversification strategy emerges from the idea that news sentiment is correlated with stock returns. Out–of–sample results indicate that such strategy dominates the benchmark index portfolio.

1.4 Essay three

HMMs have been used to analyse and predict time series phenomena. Recent work has exploited the potential of the HMM to analyse the stock

market and predict the financial market. Compared with the successful applications of HMMs in engineering, applications of HMMs in finance are in doubt. One of the main reasons for this is that most existing applications of HMM in finance use stock returns or stock prices as observations, assuming they are independent accordance with the requirements of the HMM model. However, it is apparent that prices or returns on day 1 and on the following day are not, in actual fact, independent.

In Essay three, we propose a new model (CMHMM), in which the observation is affected by a Markov model and an HMM model. The three fundamental questions of CMHMM are discussed. Further, the application of the CMHMM, in which the news sentiment as one observation and the stock return as the other observation is analysed. The empirical results of the trading strategy based on the CMHMM show the potential applications of the proposed model.

1.5 Contributions of this thesis

This thesis contributes to the literature in a number of ways. First, it extends the literature on financial applications of nonlinear models. The ANN model, the kNN algorithm and the HMM are widely used by many applications/systems in engineering, but the use of these models in finance is still under development.

This research expands the application of the ANN in finance. Most existing research on the use of ANNs in finance employs only the past stock price to predict the future direction of stock price movement. Our ANN trading

strategies are based on the information provided in news, which has only been available in the form of data over the past few years.

The existing research using HMM for stock price prediction utilises stock price or stock return as the observation that is in conflict with the assumption by the HMM models that the observations are independent. In this research, we consider the different levels of ‘economic state’ as hidden states. Each ‘economic state’ has a significant chance to generate different levels news sentiment and different levels stock return. We can observe stock return and news sentiment to estimate the hidden state. We therefore consider the return to be affected by the past stock return (a Markov model) and an HMM model. This CMHMM model adheres to the nature of the stock market and has better potential prediction ability.

Second, the empirical results from this dissertation contribute to the understanding of the relationship between the news and the stock market. For instance, we find that news volume is not the Granger cause of stock price change, but that news volume contributes to the portfolio variance both in–sample and out–of–sample; conversely, we find that news sentiment is the Granger cause of stock price change and that, as investors overact to news sentiment, positive sentiment contributes to portfolio return in–sample while negative news sentiment contributes to portfolio return out–of–sample.

Most existing research on testing the efficient markets hypothesis (EMH) focuses on the weak–form version. In this study, we consider news items as public information and test the semi–strong form efficient market using statistical arbitrage. The ability of our strategy to consistently beat the market is at odds with the EMH.

Chapter 2:

Essay 1: The relation between news items and stock price movement: An analysis based on artificial neural networks

2.1 Introduction

The efficient market hypothesis states that price movements are extremely efficient in reflecting information flows (Fama, 1970). Some studies have shown that stock prices are related to news events such as earnings announcements (Skinner, 1994), political events (Kim and Mei, 2001) and corporate takeovers (Pound and Zeckhauser, 1990), while others have failed to find convincing evidence to relate price changes to news (Joulin et al., 2008). The aim of this study is to explore the relationship between news items and stock price movement.

We first investigate the Granger causality (Granger, 1969, Granger, 1988) between news items and stock returns. Our results show that stock price change is the Granger cause of news volume and news sentiment; conversely, news volume is not the Granger cause of stock price change, whereas news sentiment is.

Moreover, we test the semi–strong form of the efficient market hypothesis using statistical arbitrage. Behavioral finance believes that information plays a significant role in human decision making and affects stock price movement. According to the EMH, it is impossible to 'beat the market' as market prices reflect all relevant information. We consider a broad range of news releases in the stock market, build ANN trading strategies with news sentiment as inputs, and perform out–of–sample forecasting. The news releases are extracted from

the unique RavenPack News Analytics database that monitors over 1,000 types and corporate actions. We find that the ANN strategies consistently outperform the passive buy–and–hold strategy; this finding is apparently at odds with the notion of the efficient market hypothesis.

The next step is to consider a particular company and test the potential profitability of the ANN strategies. Using news sentiment analytics from RavenPack Dow Jones News Analytics, we develop an ANN model to predict the stock price movements of Google Inc. (NASDAQ: GOOG) and test its potential profitability with out–of–sample prediction.

The remainder of this chapter is organised as follows. In the second section, we discuss the Granger causality test for news and stock returns. In Section 2.3, we test the semi–strong form of the efficient market hypothesis using statistical arbitrage. The empirical results from ANNs predicting the stock price movements of Google Inc. are discussed in Section 2.4. The final section concludes this chapter.

2.2 The Granger causality test for news items and stock returns

2.2.1 Theoretical background of Granger causality

In multivariate time series analysis, we often need to determine statistical causal relations between different time series. Granger causality was first proposed by Granger (Granger, 1969) in 1969 to meet this requirement. The causality test is a technique for determining whether there is an improvement in the predictability of a series when incorporating of the past of a second series, by comparison with the predictability based solely on the past of the first series. The Granger causality test is widely used to check the relationship

between different time series, such as the relationship between economic growth and energy consumption (Asafu-Adjaye, 2000, Chiou-Wei et al., 2008), the relationship between economic growth and defence spending (Joerding, 1986), the relationship between foreign direct investment and pollution (Hoffmann et al., 2005, Lee, 2009), and the relationship between foreign trade and economic growth (Awokuse, 2007, Ho et al., 2015).

Granger (1969) defines the causality for two scalar–valued, stationary, and ergodic time series $\{X_t\}$ and $\{Y_t\}$ using a simple model:

$$X_t = \sum_{j=1}^{m} a_j X_{t-j} + \sum_{j=1}^{m} b_j Y_{y-j} + \varepsilon_t \qquad (2\text{-}1)$$

$$Y_t = \sum_{j=1}^{m} c_j X_{t-j} + \sum_{j=1}^{m} d_j Y_{y-j} + \eta_t \qquad (2\text{-}2)$$

Here ε_t, and η_t are two uncorrelated white–noise series.

If some b_j is not zero, the knowledge of past Y values helps to predict current and future X values, and Y is said to Granger cause X. Similarly, X is said to Granger cause Y if some c_j is not zero. Linear least squares predictors are used when implementing this test.

Previous studies have used the Granger causality test to explore the factors causing stock price changes. For instance, Hiemstra and Jones (1994) show unidirectional Granger causality from stock returns and percentage volume changes. Ibrahim (1999) investigates the dynamic interactions between seven macroeconomic variables and the stock prices for an emerging market: Malaysia. The results show cointegration between the stock prices and three macroeconomic variables—consumer prices, credit aggregates and

official reserves. Granger et al. (2000) test the appropriate Granger relations between stock prices and exchange rates using recent Asian flu data, revealing different conclusions for different countries. Ray (2012) reports that bi–directional causality exists between stock price and foreign exchange reserve; stock price and money supply; stock price and crude oil price; and stock price and whole price index.

In this research, we want to investigate the Granger causality between news and stock returns, that is to determine whether the phenomenon of news sentiment or news volume series is significant in forecasting stock returns series (or vice versa).

2.2.2 Data description

The stock price data that we use are daily closing prices (from the year 2004 to 2012) of the Dow Jones Price Index. We compute stock returns as 100 times the first difference of the natural logarithm of the daily stock price, that is, $100 * \ln(P_t/P_{t-1})$ and obtain the stock price return series $\{\text{Stock\_return}_t\}$.

The news data used in this study are provided by RavenPack Inc., a leading provider of news analytic data (see the Appendix for further details). For every news item, there are several variables that quantify the content and form of the article. For example, the ‘relevance’ score, ranging from 0 to 100, indicates how strongly an entity is related to the underlying news story and a score of 100 indicates that the article is highly relevant. For a news story with a relevance score of 100, the ‘ENS—Event Novelty Score’ represents the level of novelty of the story. Thus, the first story reporting a categorised event receives a novel score of 100. The novelty scores of subsequent stories about

the same event follow a decay function (i.e. 100, 75, 56, 42, 32, 24, 18, 13, 10, 8, 6, 4, 3, 2, 2, 1, 1, 1, 1, 0 …). The ‘ESS—Event Sentiment Score’ represents the news sentiment for a given entity, ranging from 0 to 100, where 0 indicates extremely negative news, 50 indicates neutral news, and 100 indicates extremely positive news.

Emerging from the years spanning 2004 to 2012, there are 20,354,107 news articles in the RavenPack database. Among them, 856,071 (4.21%) have a relevance score of 100 and 342,098 (1.68%) have an event novelty score of 100. Of the news articles with a novelty score of 100, the numbers of positive news articles, negative news articles and neutral news articles are 151,309 (44.23%), 150,458 (43.98%) and 40,331(11.79%) respectively.

In this study, we seek to explore the relationship between stock return and news volume and the relationship between stock returns and news sentiment. To do so, we calculate the number of news items in a trading day and obtain the time series $\{\mathrm{News\_number_t}\}$. The news sentiment series $\{\mathrm{News\_ESS_t}\}$ are calculated as $\sum(\mathrm{ESS}-50)$ for a trading day, that is, the sum of the ESS minus 50. After this, we examine whether there exists any causal linkage between stock prices and news by conducting the Granger causality test.

2.2.3 Empirical results

In statistics, a unit root test seeks to ascertain whether a time series variable is non–stationary, as many economic and financial time series exhibit trending behavior or non–stationarity in the mean. If the data have a unit root, then some form of trend removal is required. In our research, we first conduct

Dickey–Fuller tests (Dickey and Fuller, 1979) to examine whether there is a unit root present in the three time series $\{\text{Stock\_return}_t\}$,$\{\text{News\_number}_t\}$ and $\{\text{News\_ESS}_t\}$. The Dickey–Fuller test is one of the most commonly used root tests. Our results indicate that the null hypothesis of a unit root is rejected for all three series.

Table 2.1 reports the results of our Granger causality tests for stock price returns, news volume and news sentiment. It shows that at 5% significance level, we reject the null and conclude that there is evidence to suggest stock return change is the Granger cause of news volume and news sentiment; for the news volume and stock price return, we fail to reject the null and conclude new volume is not the Granger cause of stock price return; for the news sentiment and stock price change, we reject the null and conclude news sentiment is the Granger cause of stock price change.

Table 2.1 The results of Granger causality tests

Null hypothesis	**Significance level**	**Results**
Stock return does not Granger cause news volume	0.05	Reject
	0.01	Fail to reject
Stock return does not Granger cause news sentiment	0.05	Reject
	0.01	Fail to reject
News volume does not Granger cause stock return	0.05	Fail to reject
	0.01	Fail to reject
News sentiment does not Granger cause stock return	0.05	Reject
	0.01	Reject

In the following sections, we try to predict the stock price movement using news items and past stock prices. From Table 2.1, we know that news volume is not the Granger cause of stock price return but that news sentiment is the Granger cause of stock price change. In the following sections, we do not

consider news volume, but we include news sentiment in our model for predicting stock price movement.

2.3 Test the semi–strong form of the efficient market hypothesis using ANN

Behavioral finance shows that information plays a significant role in human decision making and that financial decisions are significantly driven by emotion and mood (Nofsinger, 2005). However, the EMH proposed by Fama (1970) implies that there is no way for investors to consistently achieve superior rates of return. Fama (1970) further classifies EMH into three forms: 1) weak–form efficiency, where the information set is limited to the information contained in the past price history of the market; 2) semi–strong form efficiency, where the information set is all information that is publicly available; 3) strong–form efficiency, where the information set comprises all available public and private information available. Fama (1991) propose changes the categories: the first category covers the more general area of tests for return predictability; the second and third categories only are changed in title, not coverage, while 'semi-strong form tests' is replaced by 'event studies', and 'strong form tests' is replaced by 'tests for private information'. In this chapter, we still follow the definition of Fama (1970).

Testing the EMH is an area of enormous interest in the literature of asset pricing and investments. However, most existing studies test only weak–form efficiency and conflicting results are reported (Yen and Lee, 2008).

Empirical results from some researchers support the weak–form of efficiency. For instance, Chan et al. (1997) examine the relationships among

stock prices in 18 national stock markets for the period spanning 1961 to 1992; the results of their unit root tests suggest that the world equity markets are weak–form efficient. Aga and Kocaman (2008) test weak–form of the efficiency of the index in Istanbul Stock Exchange, concluding that it has a weak form of efficiency.

Some studies find mixed evidence on the efficient–market hypothesis. Borges (2010) discovers mixed evidence on the efficient market hypothesis using the stock market indexes of the UK, France, Germany, Spain, Greece and Portugal, from January 1993 to December 2007. The hypothesis proves valid for Portugal, Greece, France and the UK; however, the tests for Germany and Spain do not allow the rejection of the EMH. Mlambo and Biekpe (2007) also find mixed evidence for 10 African stock markets by using the runs test methodology.

Some researchers challenge the efficient market hypothesis. Lee et al. (2010) employ a panel data stationarity test that incorporates multiple structural breaks for the stock price series of 32 developed and 26 developing countries. Their results are inconsistent with the efficient market hypothesis. Further, Nisar and Hanif (2012) examine the weak form of efficient–market hypothesis in the four major stock exchanges of South Asia: India, Pakistan, Bangladesh and Sri Lanka. They apply four statistical tests–including the runs test, serial correlation, unit root and the variance ratio test for the historical index on a monthly, weekly and daily basis for a period of 14 years (from 1997 to 2011). Their findings suggest that none of the four major stock markets of South Asia follows the random–walk and hence, that none of these markets is the weak–form of efficient.

In this strand of literature, the papers can be further subdivided into two major groups. The first group of studies tests the predictability of security returns on the basis of past price changes (Gozbasi et al., 2014, Westerlund et al., 2015). More specifically, these studies employ a wide array of statistical tests to detect different types of deviations from the random walk in financial time series, such as linear serial correlations, unit roots, low–dimensional chaos, nonlinear serial dependence and long memory (Lim and Brooks, 2011). For instance, Narayan et al. (2014) test the predictability of excess stock returns for 18 emerging markets, using a range of macroeconomic and institutional factors, through a principal component analysis. Westerlund and Narayan (2013) exploit the information contained in the heteroskedasticity of the data to test EMH.

The second group of studies examines the profitability of trading strategies based on past returns, such as technical trading rules, momentum and contrarian strategies (Park and Irwin, 2007). A technical trading system consists of a set of trading rules that generate trading signals, such as long, short and out of the market signals. The profit of trading strategies is apparently at odds with the notion of the efficient market hypothesis, which implies that there is no way for investors to achieve consistently superior rates of return in an efficient market. For example, Bessembinder and Chan (1995) find that investors can earn statistically significant profits from commodity futures markets using momentum–based trading strategies. Hogan et al. (2004) empirically investigate whether momentum and value trading strategies constitute statistical arbitrage opportunities and find that these opportunities are in conflict with market efficiency.

An increasing number of researchers use ANNs in the technical analysis of the stock market. ANNs have been used to solve complicated practical problems in various fields, such as pattern recognition and medical diagnosis (Miller et al., 1992, Paliwal and Kumar, 2009). In particular, there is a burgeoning strand of literature on the applications of ANNs in economics and finance (Qi, 1996, Wong et al., 1997, Wong and Selvi, 1998, McAdam and McNelis, 2005, Schmeling, 2009). Most research using ANNs for technical analysis only considers past stock prices and volumes. For example, Leigh et al. (2002) build a neural network prediction system based on the dynamics of market price and volume. Their results support the effectiveness of the technical analysis approach. Beale et al. (2015) explore the profitability of stock trading by using a neural network model developed to assist the trading decisions of the volume adjusted moving average and the ease of movement indicator.

Most of the existing papers (Borges, 2010, Mlambo and Biekpe, 2007, Lee et al., 2010, Nisar and Hanif, 2012, Gozbasi et al., 2014) on testing the EMH have the shortcoming of not being comprehensive and robust as they only consider past price returns and/or major announcement events. Under the semi–strong form of EMH, the information set includes all publicly available information. In this research, we use the dataset from RavenPack Inc., a leading provider of news analytics data. RavenPack's News Analytics dataset systematically tracks and analyses information on more than 2,200 government organisations, 138,000 key geographical locations, 150 major currencies, 80 traded commodities and over 30,000 companies. This dataset contains almost all publicly available news and can be used to test the semi–

strong form of EMH.

In this research, we explore the financial market reaction to news items, which has been a topic of much discussion recently. For instance, Chan (2003) examines monthly returns following public news and finds a difference compared to stocks with similar returns, but with no identifiable public news. Ann et al. (2005) reveal that insider purchases (sales) are a good indication of good (bad) news and the information content of insiders trades provided that investors are able to realise returns within, at most, two months after the announcement date. Özatay et al. (2009) find that the emerging market bond index spreads respond substantially to the US. macroeconomic news and changes in the Federal Reserve's target interest rates. Bollen et al. (2011) show that measurements of collective mood states derived from large–scale Twitter feeds correlate to the value of the Dow Jones Industrial Average (DJIA) over time. Their results indicate that the accuracy of DJIA predictions can be significantly improved by the inclusion of specific public mood dimensions.

A key implication of the efficient market hypothesis is that any attempt to make profits by exploiting currently available information is futile. The market price already reflects all that can be known from the available information. The profit of trading strategies that consider all public news is apparently at odds with the notion of the semi–strong efficient market hypothesis. In this research, we build an ANN trading strategy and test the semi–strong form efficiency market hypothesis.

The remainder of this section is organised as follows. In Section 2.3.1, we introduce the ANN methodology used to generate predictions. In Section 2.3.2, we discuss the data set used in this research. The empirical results of

using ANNs to predict the direction of DJIA index movement and the trading strategy of using ANNs are discussed in Section 2.3.3. We conclude this study and discuss potential future research directions in Section 2.3.4.

2.3.1 ANN methodology

In the past decade, the availability of datasets has increased tremendously in various fields including finance. Therefore, the empirical applications of data mining techniques, such as classification, clustering and association, have become increasingly important (Liao et al., 2012). In particular, there is an emerging strand of literature on the applications of data mining techniques in the analysis of stock price movements (Paranjape-Voditel and Deshpande, 2013, Aghabozorgi and Teh, 2014, Patel et al., 2015, Li et al., 2016). This strand of literature suggests that the ANN model is fast becoming one of the leading data mining techniques in the field of stock market prediction (Kim and Han, 2000, Cao et al., 2005, Hassan et al., 2007, Guresen et al., 2011, Kara et al., 2011, Wang et al., 2011, Chang et al., 2012, Preethi and Santhi, 2012, Ticknor, 2013, de Oliveira et al., 2013). Chang et al. (2012) suggest that ANN can be employed to enhance the accuracy of stock price forecasting. de Oliveira et al. (2013) show that the ANN model is a feasible alternative to conventional techniques for predicting the trends and behavior of stocks in the Brazilian market.

The structure of the ANN model mimics the human brain and nervous system (Hill et al., 1994, Zhang et al., 1998, Bahrammirzaee, 2010). A neural network consists of a set of fundamental processing elements (called neurons) and processes information using a connectionist approach to computation.

Most neural networks contain three types of layers: input, hidden, and output (as shown in Figure 2.1). Each neuron in a hidden layer receives the input data attributes x_m from each of the neurons in an input layer and the attributes are added through applied weights w_m and converted to an output value by an activation function (u). Then, the output is passed to the neurons in the next layer, providing a feed forward path to the output layer (z).

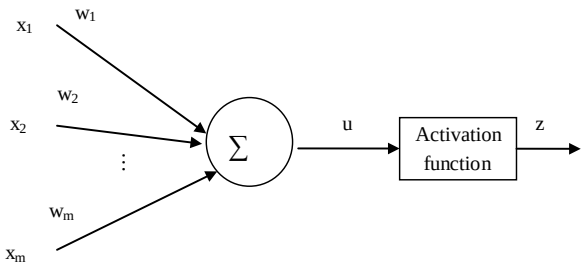

Figure 2.1 The structure of a neuron with its summation node

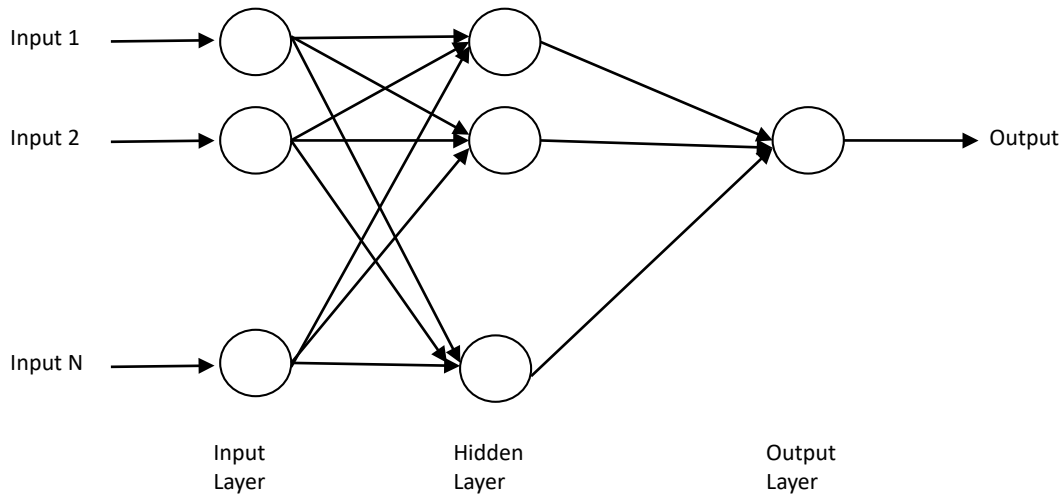

Figure 2.2 The structure of a neuron network

In general, the activation function is a nonlinear function. Activation functions that are commonly used include the threshold function, the piecewise

linear function and the sigmoid function. A neural network is an adaptive system that changes its weights based on external or internal information that flows through the network during the learning phase. These learning rules include supervised learning, unsupervised learning and reinforcement learning. In supervised learning, the network is trained through provision with input and matching output patterns.

The Probabilistic Neural Network (PNN) is one of the most widely implemented neural network topologies (Specht, 1988, Specht, 1990). PNN is devised on the basis of the classical Bayesian classifier, whose goal is to statistically minimise the risk of misclassifications. Adapting the concept of posterior probability, whose goal is to statistically minimise the risk of misclassifications a process that assumes that the probability density function of the population from which the data were drawn is known a priori—the decision rule is to classify a sample to the class with the maximum posterior probability. The PNN then uses a training set to obtain the desired statistical Bayesian information. The desired probability density function for each class is approximated using Parzen windows, a nonparametric procedure that synthesises an estimate of a probability density function by the superposition of a number of windows.

In this study, the PNN is implemented using the Neural Network Toolbox of MATLAB from Mathworks, with the network structures specified according to the default settings (Beale et al., 2015). More specifically, the PNN creates a two–layer network structure. The first layer has radial basis network neurons and calculates its weighted inputs by the distance between its weight vector and the input vector, multiplied by the bias. The second layer has competitive

transfer function neurons and calculates its weighted input using dot product weight function and its network inputs with the sum of network inputs.

2.3.2 Data

We use news items from the RavenPack News Analytics (RPNA) database as the all available public information (see the Appendix for further details). RPNA offers an analytical output for macroeconomic news release on a global basis. The database contains a unique observation for every article and includes the date and time each news article was released, a unique firm identifier, and several variables that quantify the content and form of the article.

In this research, we only consider several important fields of news items. These fields include the 'relevance' score, ranges from 0 to 100 (highly relevant) and indicates how strongly an entity is related to the underlying news story; the 'ENS–Event Novelty Score', represents its degree of novelty while the first story reporting a categorized event receives a novel score of 100; the 'ESS – Event Sentiment Score', ranging from 0 to 100, where 0 indicates extremely negative news, 50 indicates neutral news, and 100 indicates extremely positive news.

We construct the daily sentiment score (DSS) using 'ENS' and 'ESS'.

$$\mathrm{DSS_i} = \sum_{\text{all news in day i}} \mathrm{I(ENS = 100)(ESS - 50)} \qquad \text{(2-3)}$$

Determining the data frequency mainly depends on the final goal of the ANN. High–frequency data, that is, intraday data, are prone to be contaminated by noise. In this study, we use daily closing prices of the Dow Jones Price Index from 1 January 2007 to 31 December 2012. The data source

is the Thomson Reuters Tick History (TRTH) database, which is provided by the Securities Industry Research Centre of Australasia (SIRCA). We compute stock returns as 100 times the first difference of the natural logarithm of the daily stock price—that is, $100 * \ln(P_t/P_{t-1})$—and obtain stock price return series.

Behavioral finance shows that information plays a significant role in human decision making and that financial decisions are significantly driven by emotion and mood (Nofsinger, 2005). In this research, we consider stock returns and news sentiments to construct our trade strategy.

2.3.3 Empirical results

In this study we define the movement of stock prices 'up' ('down') in Day i by whether the closing Dow Jones Price Index in Day i is larger (or smaller) than the closing Dow Jones Price Index in Day i–1. We use the movement of stock prices 'up' or 'down' as our training patterns. Figure 2.3 shows our neural network classification framework for the prediction of stock price movements. The neural network model is trained using the training data and subsequently tested to measure its performance on the testing data. Basically, the process of training or learning leads to obtain the optimum neural network weights by minimising the model error, which is the difference between the actual output and the desired one. In this study, we employ data during from 1 January 2007 to 31 December 2011 as the training set and data from 1 January 2012 to 31 December 2012 as the test set. Given these preparations, this study uses stock price returns of the last three trading days (i.e.,$\text{Stock\_return}_{t-3}, \text{Stock\_return}_{t-2}, \text{Stock\_return}_{t-1}$) and the DSS of the final

trading days (i.e., $\text{Daily Sentiment Score, } \text{DSS}_{t-1}$) as input features in the PNN. In total, this approach comes to four indices meaning that there are four input nodes and one output node. We have not normalised the data because neural networks are able to recognise the high–level feature.

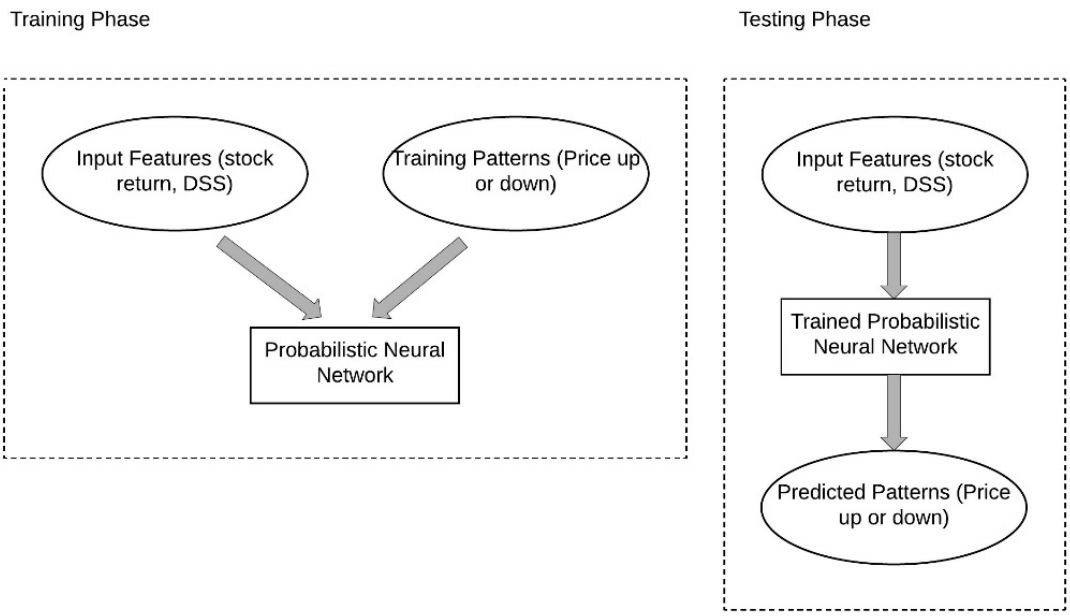

Figure 2.3 The PNN classification framework for the prediction of stock price trends

For the application of binary classification, sensitivity and specificity are used to measure the performance. In this study, we define the pattern as: 'price up' and 'price down'; then we calculate

$$\text{Sensitivity} = \frac{\text{Number of true 'up'}}{\text{Number of true 'up'} + \text{Number of false 'down'}}$$

$$\text{Specificity} = \frac{\text{Number of true 'down'}}{\text{Number of true 'down'} + \text{Number of false 'up'}}$$

$$\text{Prediction rate} = \frac{\text{Number of true 'up'} + \text{Number of true 'down'}}{\text{Number of prediction days}}$$

Table 2.2 shows the performance characteristics of our PNN prediction model. The sensitivity is 54.7% and the specificity is 58.0%. The prediction rate is 56.2%. According to the random walk directional forecast, the stock price has a fifty–fifty chance of closing higher or lower than the opening price. As can be seen, the sign predictions indicate a performance better than a random walk directional forecast.

Table 2.2 Performance Characteristics of the PNN prediction model

	Sensitivity	**Specificity**	**Prediction rate**
PNN	54.7%	58.0%	56.2%

Considering transaction costs, it is not smart to go or stay ‘long’ when the forecast return falls above zero, nor is it prudent to go or stay ‘short’ when the forecast return is below zero. The ‘long’ and ‘short’ positions are defined as buying and selling at the current price respectively. However, the trading strategy applied in this section is to go or stay ‘long’ when the forecast return is above 0.2% and to go or stay ‘short’ when the forecast return is below –0.2%. We use 0.2% to balance the transaction costs and trading frequency. We consider the estimated total return of such a strategy as:

$$\mathrm{R_{ANN}} = \sum_{t=n+1}^{n+\rho} y_t r_t \qquad (2\text{-}4)$$

Here ρ is the out–of–sample horizon and y_t the recommended position that takes the value of –1 (for a short position), +1 (for a long position) and 0 (for a hold position); r_t is the return in the Day t–that is, $r_t = 100 * \log(P_t/P_{t-1})$; P_t and P_{t-1} are closing prices of the security at Day t and Day t–1, respectively; and n is the number of training observations.

Figure 2.4 shows our neural network classification framework for the trading strategy. The training patterns are $\frac{P_t}{P_{t-1}} > 1.002,\ 1.002 \geq \frac{P_t}{P_{t-1}} \geq 0.998$ and $\frac{P_t}{P_{t-1}} < 0.998$. We employ data from 1 January 2007 to 31 December 2011 as the training set and data from 1 January 2012 to 31 December 2012 as the test set. During the test phase, we use stock price returns from the last three trading days (i.e., $\text{Stock\_return}_{t-3}, \text{Stock\_return}_{t-2}, \text{Stock\_return}_{t-1}$) and the DSS from the final trading days (i.e., $\text{Daily Sentiment Score, } \text{DSS}_{t-1}$) as input features in the PNN.

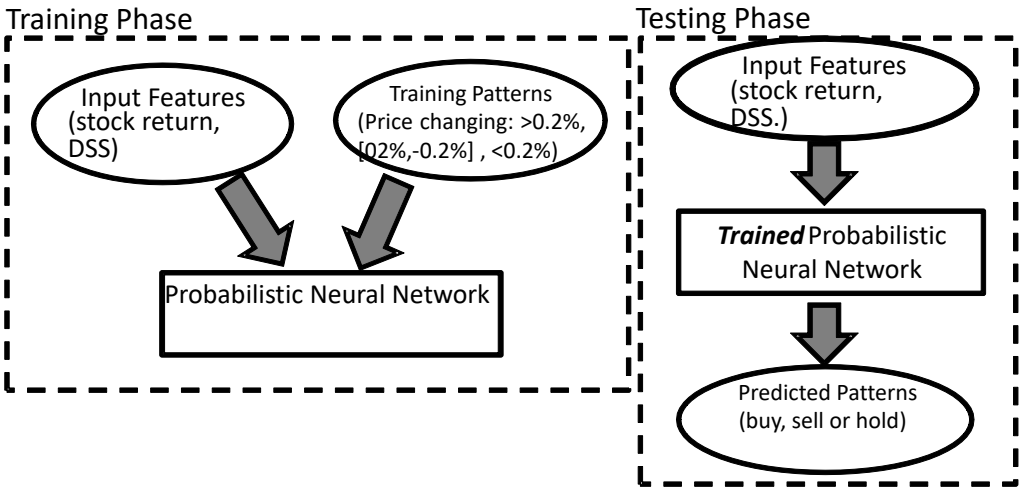

Figure 2.4 The trading strategy based on the PNN

With the efficient–market hypothesis, no mechanical trading rule would consistently outperform the buy–and–hold policy. We compare our proposed strategy with the buy–and–hold policy. The returns on a simple buy–and–hold strategy are given as follows:

$$R_b = 100 * \log(\frac{P_{t+\rho}}{P_t}) \quad (2\text{-}5)$$

Here ρ indicates the holding period, and P_t and $P_{t+\rho}$ are the prices of

securities at Time t and $t + \rho$ respectively.

In the out–of–sample testing, the total return of the proposed strategy R_{ANN} is 15.6 and the total return of the buy–and–hold strategy R_b is 7.0. In other words, using the buy–and–hold strategy, the investor can achieve a premium ($P_{t+\rho}/P_t - 1$) of 7.25%; in contrast, using the proposed strategy, the investor can access a premium of 16.89%. The trading rule based on ANNs dominates the buy–and–hold strategy.

2.3.4 Concluding remarks

Most of the existing research on testing efficient market hypothesis focuses on the weak–form version. In this study, we consider news items as public information and test the semi–strong form efficient market using statistical arbitrage. As far as we know, this study is the first study that uses news sentiment to build trade strategies. Our results show that the proposed PNN strategy outperforms the buy–and–hold strategy in terms of trading performance. The ability of our strategy to consistently beat the market is at odds with the EMH. Our models rely on powerful pattern recognition properties to produce predictions in the time series, therefore avoiding the need to specify an explicit econometric model to represent the time series.

Further, our findings suggest that news sentiment can be used to enhance the accuracy of trading strategies. Newswire message provides useful information for professional traders who can adjust their strategies proactively in response to changes in news flows and sentiment.

The key factor in using statistical arbitrage to test efficient–market hypothesis is the profitable trading strategy. However, much work is required

to improve the prediction accuracy of our model. One possible direction for future research is the use of high–frequency data of stock prices and news and the stock market volatility in the forecast model.

2.4 Predicting the stock price movement of Google Inc.

Technical analysis is the study of past price movements with the aim of forecasting potential future price movements. Market participants who use technical analysis often exploit primary market data, such as historical prices, volume and trends, to develop trading rules, models and even technical trading systems. These systems comprise a set of trading strategies and rules that generate trading signals, for example, buy and sell signals, in the market. Several studies (Bessembinder and Chan, 1995, Fernandez-Rodrıguez et al., 2000, Hsu and Kuan, 2005, Park and Irwin, 2007, Han et al., 2013) examine the profitability of these trading strategies, which include moving average, momentum and contrarian strategies. In particular, Park and Irwin (2007) suggest that out of 95 modern studies on technical trading strategies, 56 of them provide statistically significant evidence that technical analysis generates positive results. Han et al. (2013) demonstrate that a relatively straightforward application of a moving average timing strategy outperforms the passive buy–and–hold strategy. Bessembinder and Chan (1995) suggest that technical trading rules have varying degrees of success across different international stock markets; in general, these rules tend to be more successful in the emerging markets. Fernandez-Rodrıguez et al. (2000) examine the profitability of a simple technical trading rule based on the ANNs and conclude that the ANN trading rule is mostly superior to a passive buy–and–hold trading strategy

during ‘bear’ market and ‘stable’ market episodes.

Most existing research on technical trading rules and strategies focuses on objective and unambiguous rules based on historical market information without considering investor sentiment. Recent research in behavioural finance apparently indicates that news sentiment is significantly related to stock price movements (Neal and Wheatley, 1998, Antweiler and Frank, 2004, Schmeling, 2009, Lux, 2011, Chung et al., 2012, Wang et al., 2013). For instance, Antweiler and Frank (2004) suggest that internet messages have a significant impact on stock returns and disagreement among posted messages is associated with increased trading volumes. Schmeling (2009) finds that sentiment negatively forecasts aggregate stock market returns on average across countries. Moreover, Schmeling (2009) suggests that the effect of sentiment on stock returns is higher for countries with less market integrity that are more susceptible to market overreaction and herding. Wang et al. (2013) provide evidence that, while news volume does not Granger cause stock price change, news sentiment does Granger cause stock price change. In general, these papers suggest that the effect of sentiment on stock markets cannot be ignored.

In this research, we combine a trading strategy based on the ANN model with news sentiment analysis to build our ANN model of predicting the stock price movements of Google Inc. (NASDAQ: GOOG). GOOG is an American public corporation specialising on Internet–related services and products that enhance the ways people connect with information. Its primary source of revenue comes from delivering online advertising that is relevant to consumers and cost–effective for advertisers (Google Inc., 2015). Founded by Larry Page

and Sergey Brin as a privately held company in 1998, GOOG became a public corporation after its initial public offering (IPO) on 19 August 2004. Over the past decade, its shares have grown by more than 1,500%. As of 31 December 2014, Google had 53,600 full–time employees. Its current range of services includes web search research, email, mapping, office productivity and video sharing services. We focus on GOOG for the following reasons: first, as a major stock on NASDAQ, GOOG is one of the few that has relatively straightforward transaction data because it is a non–dividend–paying stock. As noted on Google's Investor Relations website, Google has 'never declared or paid a cash dividend nor do we expect to pay any cash dividends in the foreseeable future (Google Inc., 2015). Second, since its IPO, GOOG is considered one of the best performers in the stock market, as its stock price has increased by more than 15 times over the past decade. Third, GOOG has a very high volume of outstanding shares (over 300 million with an average daily trading volume of 2.4 million) and a high stock price (over $600 in September 2015), making it unlikely to be the subject of price manipulation (Google Inc., 2015). Fourth, as a frequently traded share with a large market capitalisation exceeding US$400 billion, news directly related to GOOG is frequently reported in various major media outlets. These news releases are a rich source of data for examining the effect of news sentiment on GOOG's price movements. To quantify the sentiment associated with each news release, we use the dataset obtained from the RavenPack News Analytics Global Equities editions.

The remainder of this section is organised as follows. In Section 2.4.1, we introduce the research background. In Section 2.4.2, we discuss the

datasets used in this research. The empirical results of using ANNs to predict the stock price movements of Google Inc. are discussed in Section 2.4.3. The final section concludes this part of the thesis.

2.4.1 Research background

The stock market is essentially dynamic, nonlinear, complicated nonparametric and hard to predict. The successful prediction of a stock's future price could yield a significant profit. A large amount of research has been published using different techniques to predict the stock market. Most studies predict the movement of stock market indexes, such as the DJIA index (Quah, 2008, Cervelló-Royo et al., 2015), the NYSE composite index (Leigh et al., 2002), NASDAQ Stock Exchange index (Guresen et al., 2011), and the stock market indexes of undeveloped countries (Kara et al., 2011, de Oliveira et al., 2013).

There are few studies that predict the movement of individual companies and, further, these studies consider only past prices when predicting. For instance, White (1988) uses neural network modelling and learning techniques to search for and decode nonlinear regularities in the prediction of IBM common stock daily returns; and Hui and Chan (2014) construct two trading strategies for 12 constituent stocks of the Hang Seng Index.

The value of ANN modelling techniques in performing complicated pattern recognition and nonlinear forecasting has been shown by their applications in different domains. In this section, we apply the ANN models to predict the stock price movement of Google Inc. by considering stock returns and news sentiment.

2.4.2 Data

In this research, we use the daily opening and closing prices of GOOG from 1 January 2013, to 31 December 2014 as the training set and data from 1 January 2015 to 30 June 2015 as the test set to test the predictive accuracy of the proposed PNN method. The prices are obtained from SIRCA's TRTH database. Figure 2.6 shows the daily closing prices of GOOG from 1 January 2013, to 30 June 2015. In 2014, the price experienced a surge, which was a result of the Google two–for–one stock split on 3 April 2014. As a result of the stock split, GOOG's shareholders received two shares (Class A and Class C) for every one share that they owned. The main difference between these two classes is that Class A confers voting rights whereas Class C does not.

To compute the returns of GOOG, we calculate the difference between the natural logarithm of the daily opening and closing stock prices and multiply the difference by 100—that is, $100 * \ln(\mathrm{P_{iopen}}/\mathrm{P_{iclose}})$. $\mathrm{P_{iopen}}$ is the opening price of the GOOG and $\mathrm{P_{iclose}}$ is the closing price of the GOOG in Day i. As we use the daily opening and closing stock price to calculate the stock returns, the stock split of GOOG does not affect our estimation results.

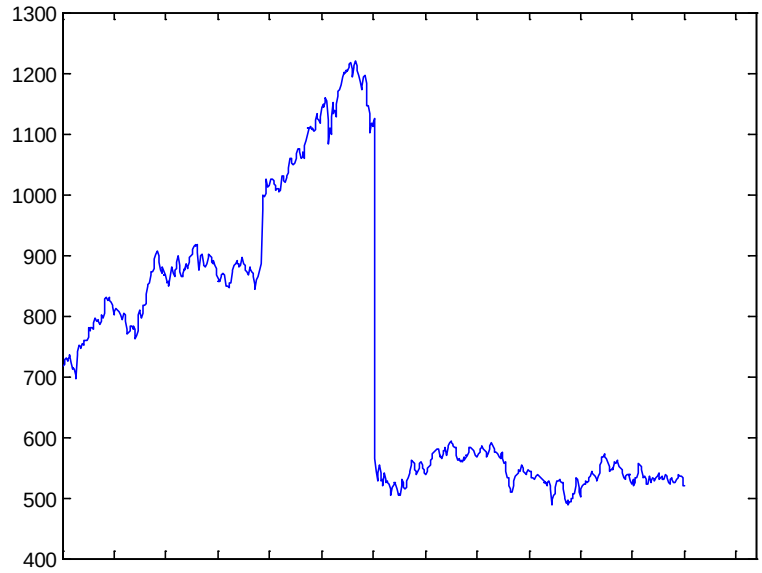

Figure 2.5 Daily closing prices of the NASDAQ: GOOG (Jan 2013–Jun 2015)

The news dataset for GOOG is obtained from RavenPack News Analytics (see the Appendix for details), which provides sentiment analysis for the news articles relevant to GOOG. For each news article, RavenPack provides the following key information: the date and time each news article is released, a unique firm identifier, and several variables that measure the relevance, content, sentiment and form of the article. In this research, we consider the 'Relevance' score and the ESS.

We construct the Daily Sentiment Score (DSS) for GOOG using the relevance score and ESS based on the formula provided below. The period that we use to calculate the DSS on Day i–1 is the 24–hour period before the market opens on day i:

$$\mathrm{DSS}_{\mathrm{i}-1} = \sum_{\text{all news about the given firm in 24 hours before market open in the day i}} \mathrm{I}(\mathrm{Relevance} = 100)(\mathrm{ESS} - 50)$$

Figure 2.6 shows the DSS for GOOG from 1 January 2013 to 30 June 2015.

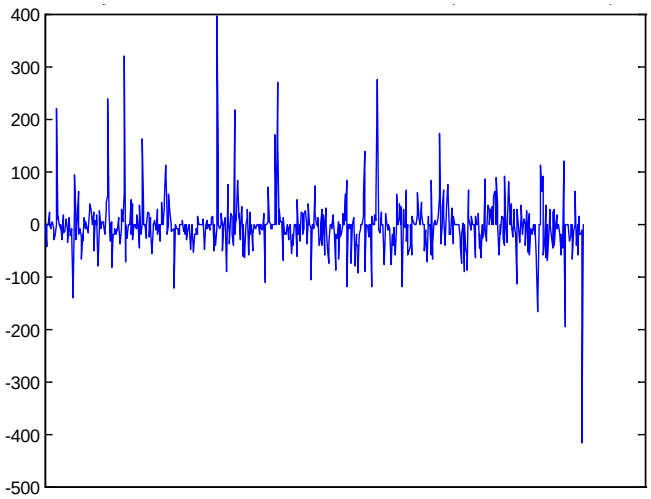

Figure 2.6 DSS of the NASDAQ: GOOG (Jan 2013–Jun 2015)

2.4.3 Empirical results

In this research, the stock price movement is ‘up’ (or ‘down’) in Day i if the closing price of GOOG on Day i is higher (or lower) than the opening price of GOOG on Day i. We use the ‘up’ and ‘down’ movements of the stock prices as our training patterns. The PNN model is trained on the training data and subsequently tested to assess its performance on the testing data. The process of training or learning helps us obtain the optimum ANN weights by minimising the model error, or the difference between the actual output and the desired one. Given these sets, this study uses stock price returns from the last three trading days (i.e., $\text{Stock\_return}_{t-3}, \text{Stock\_return}_{t-2}, \text{Stock\_return}_{t-1}$) and DSS from the final trading day (i.e., DSS_{t-1}) as input features for the PNN model. Following this approach, four indices have been obtained. In other words, there are four input nodes and one output node. Table 2.3 shows the

basic statistics of these inputs (the close–to–open return and the DSS of GOOG).

Table 2.3 The basic statistics of the close–to–open return and the DSS of GOOG (Jan 2013–Jun 2015)

	Mean	**Highest**	**Lowest**
Close–to–open return	–0.03%	4.03%	–5.48%
DSS	–2.39	396	–418

For the application of the binary classification in the PNN model, sensitivity and specificity are used to assess the performance of the model. In this research, we define the patterns as 'price up' and 'price down'; then we calculate the following variables:

$$\text{Sensitivity} = \frac{\text{Number of true 'up'}}{\text{Number of true 'up'} + \text{Number of false 'down'}}$$

$$\text{Specificity} = \frac{\text{Number of true 'down'}}{\text{Number of true 'down'} + \text{Number of false 'up'}}$$

$$\text{Prediction rate} = \frac{\text{Number of true 'up'} + \text{Number of true 'down'}}{\text{Number of prediction days}}$$

Table 2.4 shows the performance characteristics of our PNN prediction method. The sensitivity is 52.83% and the specificity is 55.71%. The prediction rate is 54.47%. The sign predictions indicate that the PNN method can perform better than a random walk directional forecast.

Table 2.4 Performance characteristics of the PNN prediction model

	Sensitivity	**Specificity**	**Prediction rate**
PNN	52.83%	55.71%	54.47%

2.4.4 Concluding remarks

Many papers on trading strategies build trading rules based on historical data such as stock price and volume. In this paper, we use the sentiment scores of news articles related to GOOG to develop an ANN model to predict its stock price movements. More specifically, by defining an ‘up’ (or ‘down’) movement on Day i as the day’s closing price being higher (or lower) than its opening price, our empirical results provide better predictive accuracy than a random walk directional forecast. Our model provides a potentially profitable trading strategy with the following rules: if the model predicts an ‘up’ movement, we should buy the stock at the stock market opening and sell the stock at the stock market close; in contrast, if the model predicts a ‘down’ movement, we should sell the stock at the stock market open and buy the stock at the stock market close.

2.5 Conclusion

In Essay one, we investigate the relationship between news and stock price changes. Our results show that stock price change is the Granger cause of news volume and news sentiment; news volume is not the Granger cause of stock price change while news sentiment is the Granger cause of stock price change. We contribute to the literature concerning efficient market hypothesis testing by our unique focus on the semi–strong form efficient market hypothesis, as most existing research on testing efficient market hypothesis focuses on the weak–form version.

Chapter 3:

Essay 2: Can news volume and news sentiment contribute to portfolio selection?

3.1 Introduction

Investors want to build a robust portfolio strategy to seek profits while avoiding the potential risks of loss. Markowitz's mean–variance model, which is the start of modern portfolio theory (MPT), was introduced more than 60 years ago and is still considered one of the most popular approaches to portfolio optimization. Markowitz's mean–variance model (1952) derived the optimal rule for allocating wealth across risky assets in a static setting when investors care only about the mean and variance of a portfolio's return. This investment theory is based on the idea that risk–averse investors can construct portfolios to optimise or maximise expected return by considering a given level of market risk and emphasising that risk is an inherent part of higher reward. When an investor constructs a portfolio, he or she has to consider how each security cooperates with all other securities. Markowitz's mean–variance optimization method suggests that it is possible to construct an 'efficient frontier' of optimal portfolios, offering the maximum possible expected return for a given level of risk.

Some researchers doubt the efficiency of Markowitz's mean–variance portfolio optimization method. For instance, the empirical study of DeMiguel et al. (2009) evaluates the mean–variance portfolio method across seven empirical datasets and finds that the mean–variance portfolio method leads to poor out–of–sample performances, no better than the 1/N rule in terms of

Sharpe ratio, certainty–equivalent return or turnover.

The big data era brings about huge challenges for portfolio diversification. Researchers consider using big data, such as stock messages (Antweiler and Frank, 2004) or information of searched items on Google Trends (Kristoufek, 2013) to help portfolio diversification.

In this chapter, we expand the use of information from news volume and news sentiments to portfolio diversification. We discuss the possibility of news volume and news sentiments contributing to portfolios by assessing the performance of portfolios that are built based on news volume and news sentiments. Our empirical analytics use the time series provided by the news analytics data from Raven Pack. Further, we propose a novel approach to portfolio diversification based on the k–Nearest Neighbors (kNN) algorithm. The diversification strategy arises from the idea that news sentiment is correlated with stock returns.

The remainder of the chapter is organised as follows. In the second section, we introduce the research background. In Section 3.3, we discuss the data set used in this research. The empirical results are reported in Section 3.4 and the robust checks are discussed in Section 3.5. The final section concludes this chapter.

3.2 Research Background

3.2.1 Modern Portfolio Theory

MPT begins with the path–breaking work of Markowitz (1952) who derived the optimal rule for allocating wealth across risky assets in a static setting when investors care only about the mean and variance of a portfolio's

return. In 1990, Harry Markowitz shared a Nobel Memorial Prize in Economic Sciences for his 'pioneering work in the theory of financial economics'. Markowitz's mean–variance optimization method, which can be traced back to his paper 'Portfolio Selection' in the Journal of Finance in 1952 and his book 'Portfolio Selection: Efficient Diversification' in 1959, suggests that it is possible to construct an 'efficient frontier' of optimal portfolios, offering the maximum possible expected return for a given level of risk.

Markowitz's mean–variance optimization method can be explained as follows. Consider a portfolio with n different assets where asset number i will give the return $\mathrm{R_i}$. Note that μ_i and σ_i^2 are the corresponding mean and variance and that $\sigma_{i,j}$ is the covariance between $\mathrm{R_i}$ and $\mathrm{R_j}$. The investor is a 'rational man' and he or she always chooses the portfolio with the smallest variance of return (i.e., the smallest risk) if the expected returns are the same, or the portfolio with the highest expected return if the variance levels are equal. For a portfolio, if the investor invests $\mathrm{x_i}$ of the value of the portfolio in asset $i, (1 > \mathrm{x_i} > 0, i = 1, 2, \dots, \mathrm{n}$ and $\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x_i} = 1)$, then the expected return of the whole portfolio R is $\mu = \mathrm{E(R)} = \mathrm{E}(\sum \mathrm{x_i} * \mathrm{R_i})$. The variance of the entire portfolio is $\sigma^2 = \mathrm{V(R)} = \mathrm{V}(\sum \mathrm{x_i} * \mathrm{R_i})$. For different choices of $x_1, x_2, \dots x_n$, the investor will receive different combinations of μ and σ^2. Those (σ^2, μ) with minimum σ^2 for a given μ or more and with maximum μ for a given σ^2 or less are called efficient frontiers, which shows that by investing in more than one asset and choosing the right combination of assets, an investor can benefit from diversification and particularly from a reduction in portfolio risk.

There are a number of critical underlying assumptions in MPT about the behavior of individuals (Beyhaghi and Hawley, 2013). These assumptions

include: 1) the investors are rational; 2) investors are risk averse and make decisions based on the axioms of expected utility theorem; 3) investors always prefer portfolios with higher expected returns if variances of the returns are equal; 4) investors are price takers who cannot affect a security price; 5) investors know the expected return of each asset in their portfolios.

Following Markowitz's work, new contents have been introduced to extend the MPT framework. The Capital Asset Pricing Model (CAPM) proposed by Sharpe (1964) takes into consideration the equilibrium asset–pricing consequences of investors' individually rational actions and provides a foundation for an asset pricing model. The CAPM model suggests that an efficient portfolio is actually a linear combination of the market portfolio and the risk–free asset. Instead of considering a single risk factor, Ross (1976) proposed Arbitrage Pricing Theory (APT), which is a generalisation of CAPM. In APT, assets returns are driven by multiple risk factors.

3.2.2 The development of Markowitz's mean–variance approach

After the pioneering work of Markowitz (1952), researchers develop mean–variance approach from different directions (Elton and Gruber, 1997). Some scholars (Lee, 1977, Konno et al., 1993, Briec et al., 2007) propose alternative portfolio theories that include additional moments such as skewness or more realistic descriptions of the distribution of returns. For instance, Briec et al. (2007) propose a nonparametric efficiency measurement approach for the static portfolio selection problem in mean–variance–skewness space.

Second, mean–variance portfolio theory was developed to find the

optimum portfolio by considering return distributions over a single period. Therefore, the other research direction of MPT concerns how the single–period problem should be modified when investors consider a multi–period investment. This problem has been analyzed under various assumptions (Celikyurt and Özekici, 2007, Gülpınar and Rustem, 2007, Calafiore, 2008, Takano and Gotoh, 2014).

Third, some researchers have sought to improve Markowitz's mean–variance approach by reducing statistical errors in the sample mean and covariance matrix. For example, Lai et al. (2011) propose a new approach to resolve the 'Markowitz optimization enigma'–a phrase that describes portfolios that may perform poorly because the means and covariances of the underlying assets are unknown and have to be estimated from historical data. Jobson and Korkie (1980) examine the sampling properties of the conventional estimators for the parameters of an efficient portfolio.

Fourth, some researchers discuss optimal portfolios under constraint (Snell and Tonks, 1998, Aktas et al., 2008, Bera and Park, 2008, Landsman, 2010). For instance, Snell and Tonks (1998) discuss efficient frontiers and optimal investment strategies for the dynamic mean–variance portfolio selection problem under the constraint of a higher borrowing rate. Aktas et al. (2008) propose a tail mean–variance approach, based on tail condition expectations and tail variance as a measure for the optimal portfolio selection. Bera and Park (2008) propose the use of a cross–entropy measure as the objective function with side conditions produced by the mean and variance–covariance matrix of the resampled asset returns. Landsman (2010) proposes the tail mean–variance approach, based on the tail condition expectation and

tail variance as a measure for the optimal portfolio selection.

3.2.3 Criticisms of Markowitz's mean–variance approach

Some researchers doubt the efficiency of Markowitz's mean–variance portfolio optimization method. Contrary to the notion of diversification, Bera and Park (2008) find that using Markowitz's mean–variance portfolio optimization method leads to portfolios that are highly concentrated on a few assets and result in poor out–of–sample performances. Scherer (2002) show that when using the mean–variance portfolio optimization method, small changes in inputs can give rise to large changes in the portfolio.

Additionally, using Markowitz's mean–variance portfolio optimization theory, investors must assume that the means and covariance of the underlying asset returns are known, whereas in practice, they are unknown. Normally, they are estimated using historical data and led to portfolios that may perform poorly. As stated above, the empirical study of DeMiguel et al. (2009) evaluates the mean–variance portfolio method across seven empirical datasets, finding that it leads to poor out–of–sample performances, no better than the 1/N rule in terms of Sharpe ratio, certainty–equivalent return, or turnover.

On the whole, using the mean–variance optimization method, investors solely base on expected return and risk to make decisions. These expectations are derived from historical returns. Their optimal asset allocations are highly sensitive to small changes in inputs and may not be well diversified.

3.2.4 Big data opportunity

Recently, data harvesting has increased on a large scale and across various fields. The concept of ‘Big data’ not only relates to the storage of and access to data but also to the way in which data are understood and exploited. Researchers seek to comprehend the relationship between the news, the stock return and market volatility. For instance, Antweiler and Frank (2004) study the effect of more than 1.5 million messages posted on *Yahoo! Finance* and *Raging Bull* about the 45 companies in the DJIA and the Dow Jones Internet Index. They find that stock messages help predict market volatility and that the effect of these messages on stock returns is statistically significant but economically small. Alanyali et al. (2013) exploit a large corpus of daily print issues of the *Financial Times*—from 2 January 2007 until 31 December 2012—to quantify the relationship between decisions made in financial markets and developments in financial news. They find a positive correlation between the daily number of mentions of a company in the *Financial Times* and the daily transaction volume of that company’s stock, both on the day before the news is released and on the same day as the news is released.

An increasing number of scholars are considering building portfolios or improving portfolio optimization methods according to big data or using big data analytic techniques. To estimate portfolio risk, Mitra et al. (2009) present a tractable method of including both option implied volatility and quantified news. Kristoufek (2013) discusses an approach to portfolio diversification using the information of searched items on Google Trends. In his research, the popular stocks are penalised by assigning them lower portfolio weights and

bring forward the less popular or peripheral stocks to decrease the total riskiness of the portfolio. His results indicate that such a strategy dominates both the benchmark index and the uniformly weighted portfolio both in–sample and out–of–sample. Gillam et al. (2015) propose a measure of abnormal news volume that controls for the size of the firm and the analyst attention that it receives and demonstrate that this measure enhances the predictive power of the global stock selection model using information coefficients, Boolean signals and efficient frontiers. Creamer (2015) advocates a portfolio diversification method that outperforms the market portfolio. In his method, investors' expectations are based either on news sentiment using high–frequency data or on a combination of accounting variables, financial analyst recommendations and corporate social network indicators with quarterly data.

3.2.5 The research plan

Research in behavioural finance indicates that news sentiment is significantly related to stock price movements and financial decisions are significantly driven by mood and sentiment (Nofsinger, 2005). For instance, Tetlock (2007) quantitatively measures the interactions between the media and the stock market using daily content from a popular *Wall Street Journal* column. He finds that high media pessimism predicts downward pressure on market prices followed by a reversion to fundamentals and that unusually high or low pessimism predicts high market trading volume. Zouaoui et al. (2011) examine the influence of investor sentiment on the probability of stock market crises and find that investor sentiment increases the probability of occurrence of stock market crises within a one–year horizon. Wang et al. (2013) show

evidence that while news volume does not Granger cause stock price change, news sentiment does Granger cause stock price change. Ho and Wang (2016) develop an ANN model to predict the stock price movements of GOOG and test its potential profitability using out–of–sample prediction. In general, these papers suggest that the effect of sentiment on stock markets cannot be ignored.

In this research, we consider the utility of information from news volume and news sentiment to portfolio diversification. For the DJIA components, we assign different weights according to their weekly news volume or news sentiment to build portfolios. We follow the power–law rule proposed by Kristoufek (2013) to obtain the weights of our portfolio components. Both in– and out–of–sample are used to assess the performance of the portfolios. The former is a standard approach employed to measure the quality of portfolio optimization and the latter is more useful for evaluating the practical applicability of portfolio selection methods. The in–sample refers to the building of portfolio weights using information from the same period, while the out–of–sample refers to the building of portfolio weights at Week t using the information in Week t–1.

Further, we propose a novel approach to portfolio diversification based on the k–Nearest Neighbors (kNN) algorithm. The diversification strategy is based news sentiment is correlated with stock returns.

3.3 Data

3.3.1 Dow Jones Industrial Average index

The DJIA, also called the Dow 30 or simply the Dow, is a stock market index that was first calculated on 26 May 1896. The DJIA is the most–quoted

market indicator in newspapers, on TV and on the Internet. The DJIA comprised only 12 stocks at its beginning and expanded to 20 firms in 1920. In 1928, the industrial average was expanded to its current level of 30 firms, which, on the DJIA, have historically accounted for approximately 25% of the market value of all NYSE firms (Jones et al., 1989). The DJIA is one of the most important indexes of the NYSE and it reliably indicates basic market trends.

The index shows how 30 large publicly owned companies based in the United States have traded during a standard trading session in the stock market. The DJIA is price weighted rather than market capitalization weighted. In other words, its component weightings are affected only by changes in the stock prices. Additionally, the practice of adjusting the divisor has been initiated to mitigate the effects of stock splits and other adjustments. Figure 3.1 shows the daily closing values of the DJIA from January 2014 to June 2016.

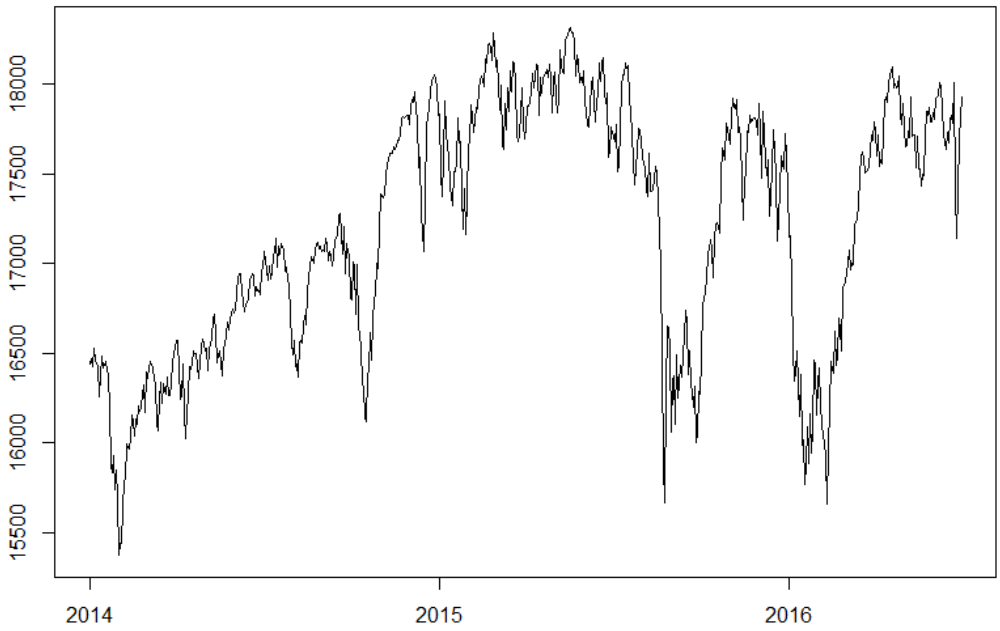

Figure 3.1 Daily closing values of the DJIA (Jan 2014–Jun 2016)

The components of the DJIA have changed many times since its inception; the reasons for these changes are: i) firm mergers and reorganisations and; ii) the achievement of a better representation of American industry (Jones et al., 1989). The most recent change to the index occurred on 19 March 2015 when Apple replaced AT&T, which had been a component of the DJIA since November 1916. Table 3.1 shows the 30 major American companies that currently comprise the DJIA.

Table 3.1 DJIA components (since 19 March 2015)

	Company Name	**Exchange**	**Symbol**
1	3M Company	NYSE	MMM
2	E.I. du Pont de Nemours & Company	NYSE	DD
3	McDonald's Corporation	NYSE	MCD
4	Exxon Mobil Corporation	NYSE	XOM

	Company Name	Exchange	Symbol
5	Merck & Co. Inc.	NYSE	MRK
6	American Express Company	NYSE	AXP
7	General Electric Company	NYSE	GE
8	Microsoft Corporation	NASDAQ	MSFT
9	Pfizer Inc.	NYSE	PFE
10	The Home Depot Inc.	NYSE	HD
11	The Procter & Gamble Company	NYSE	PG
12	The Boeing Company	NYSE	BA
13	Intel Corporation	NASDAQ	INTC
14	The Travellers' Companies Inc.	NYSE	TRV
15	Caterpillar Inc.	NYSE	CAT
16	International Business Machines Corporation	NYSE	IBM
17	United Technologies Corporation	NYSE	UTX
18	Chevron Corporation	NYSE	CVX
19	Johnson & Johnson	NYSE	JNJ
20	Verizon Communications Inc.	NYSE	VZ
21	Cisco Systems, Inc.	NASDAQ	CSCO
22	JPMorgan Chase & Co.	NYSE	JPM
23	Wal–Mart Stores Inc.	NYSE	WMT
24	The Coca–Cola Company	NYSE	KO
25	The Walt Disney Company	NYSE	DIS
26	UnitedHealth	NYSE	UNH
27	Goldman Sachs	NYSE	GS
28	Nike	NYSE	NKE
29	Visa	NYSE	V
30	Apple	NASDAQ	AAPL

Note: NYSE refers to the New York Stock Exchange. NASDAQ is the acronym for the National Association of Securities Dealers Automated Quotations, which is the second–largest exchange in the world by market capitalisation, behind only the NYSE.

The DJIA tracks only 30 large American companies; however, these companies are inclusive of all industries except utilities and transportation, creating a broad overview of the economy. In general, the DJIA is a leading indicator and is considered by many investors to represent trends in the economy. Many researchers use the DJIA in their study of the stock market.

For instance, Beneish and Gardner (1995) examine the stock market effect of changes in the composition of the DJIA. Charles and Darné (2014) explore the relations between events (e.g., financial crashes, elections, wars and monetary policies) and the consequent volatility of the DJIA index during the period from 1928 to 2013.

3.3.2 Data acquisition and pre–processing

In this research, we consider the stock price returns and news items for only two and a half years (Jan 2014–Jun 2016). We consider only 29 of the 30 DJIA index components; we do not include Apple Inc. (NASDAQ: AAPL) or AT&T Inc. (NYSE: T) in our research (Apple Inc. replaced AT&T Inc. on 19 March 2015).

For each of the 29 stocks, we construct the series of daily returns, $r_{i,j}$, defined as $r_{i,j} = \frac{p_{i,j+1} - p_{i,j}}{p_{i,j}}$. The $p_{i,j}$ is the adjusted opening price of stock i on Day j. This approach is different from those of most studies that use the closing price to calculate the daily return as we adjust our portfolio at the stock market's opening according to the news. The adjusted price is used to produce an accurate representation of the firm's equity value beyond the simple market price. The adjusted opening price considers all corporate actions, such as stock splits, dividends and distributions and rights offerings. The adjusted opening and closing price data have been obtained using *Yahoo! Finance* (https://finance.yahoo.com/).

Our raw news item data have been obtained from the RPNA database (see the Appendix for further details). The database contains unique observations for every article and includes the date and time each news article

was released, a unique firm identifier and several variables that quantify the content and form of the article. During the period from January 2014 to June 2016, there were a total of 15,476,000 news items; there were 6,265,885 in 2014, 6,103,673 in 2015 and 3,106,442 from January to June in 2016.

There are 39 fields used to describe each news item. In this research, we consider only some of these fields, such as time stamp, company name, news relevance, event sentiment and news novelty.

We ascertain the news relevant to the 29 stocks based on the field 'company name'. To analyse the effect of the news on the stock market, we define the news that happens before the stock market's opening on Day i as the news that happens on the Day $i-1$. When building the daily news volume and news sentiment series, we need to consider the market hours of the NASDAQ stock market and the NYSE, which run from 9:30am to 4:00pm on weekdays. Further, we need to consider summer daylight–savings time adjustments when pre–processing our dataset as the RPNA uses coordinated universal time (UTC) for every news item, so that 2:00am on 9 March 2014, 8 March 2015 and 13 March 2016 becomes 3:00am; and 2:00am on 2 November 2014 and 1 November 2015 becomes 1:00am.

Table 3.2 shows the basic statistics for the daily news volume of these 29 stocks. If we only consider the new news items, the average daily news volume is less than one article per stock.

Table 3.2 The basic statistics for the daily news volume of the 29 stocks (Jan 2014–Jun 2016)

		News volume (all)					News volume (only new news)				
	Symbol	**Mean**	**Median**	**Max**	**Min**	**Std. dev.**	**Mean**	**Median**	**Max**	**Min**	**Std. dev.**
1	MMM	5.58	3	123	0	11.27	0.34	0	14	0	1.38
2	DD	9.40	3	217	0	19.68	0.34	0	13	0	1.27
3	MCD	17.93	11.5	205	0	24.68	0.42	0	16	0	1.54
4	XOM	20.70	15	223	0	23.42	0.38	0	15	0	1.37
5	MRK	12.49	7	168	0	18.19	0.49	0	17	0	1.44
6	AXP	14.21	10	139	0	16.13	0.38	0	12	0	1.14
7	GE	37.29	32.5	406	0	36.29	0.95	0	15	0	1.72
8	MSFT	42.75	40	346	0	38.16	0.68	0	13	0	1.45
9	PFE	21.54	15	298	0	30.80	0.53	0	16	0	1.62
10	HD	11.76	7	244	0	20.54	0.33	0	17	0	1.52
11	PG	14.14	11	197	0	19.49	0.44	0	18	0	1.45
12	BA	30.88	24	236	0	31.19	0.66	0	19	0	1.76
13	INTC	15.36	11	144	0	19.36	0.62	0	19	0	1.52
14	TRV	4.91	4	49	0	5.24	0.23	0	6	0	0.53
15	CAT	8.80	4	215	0	19.29	0.40	0	15	0	1.53
16	IBM	22.10	18	351	0	25.95	0.71	0	14	0	1.46
17	UTX	9.93	5	198	0	16.43	0.50	0	17	0	1.69
18	CVX	16.50	11	214	0	20.19	0.35	0	15	0	1.25
19	JNJ	15.90	12	211	0	21.31	0.45	0	18	0	1.50
20	VZ	23.91	19	258	0	25.27	0.53	0	13	0	1.38
21	CSCO	15.46	13	134	0	17.20	0.49	0	13	0	1.30
22	JPM	131.46	123	1035	0	115.69	2.99	2	41	0	3.84
23	WMT	27.64	21	317	0	31.57	0.60	0	20	0	1.82
24	KO	19.95	14	265	0	24.94	0.48	0	16	0	1.39
25	DIS	20.30	16	200	0	21.56	0.39	0	11	0	1.09
26	UNH	7.81	4	184	0	16.14	0.32	0	13	0	1.14
27	GS	106.54	96	1287	0	114.98	2.13	1	27	0	3.12
28	NKE	11.32	6	175	0	15.85	0.37	0	12	0	1.25
29	V	6.02	2	140	0	12.31	0.32	0	20	0	1.33

This table presents a summary of the descriptive statistics for the daily news (all news and only new news) volume (Jan 2014–Jun 2016) for the 29 stocks used in this study. The new news is defined as ENS = 100. The summary statistics include the mean value (Mean), median value (Median), maximum (Max), minimum (Min) and standard deviation (Std. dev.).

For each news item, the ESS represents the news sentiment for a given

entity, ranging from 0 to 100, where 0 indicates extremely negative news, 50 indicates neutral news and 100 indicates extremely positive news. To easily understand the effect of the news, we use –50 to indicate extremely negative news, 0 to designate neutral news and +50 to connote extremely positive news for each item. Table 3.3 shows the basic statistics of daily total news sentiments of these 29 stocks.

Table 3.3 The basic statistics for the daily total news sentiment for the 29 stocks (Jan 2014–Jun 2016)

	Symbol	**Mean**	**Median**	**Max**	**Min**	**Std. dev.**
1	MMM	1.46	0	119	-39	10.25
2	DD	1.16	0	146	-41	10.41
3	MCD	0.39	0	148	-78	11.16
4	XOM	0.88	0	61	-49	8.49
5	MRK	2.76	0	116	-56	11.67
6	AXP	1.48	0	120	-62	10.13
7	GE	9.57	0	166	-32	21.31
8	MSFT	3.92	0	136	-41	13.22
9	PFE	3.26	0	180	-56	15.28
10	HD	1.86	0	194	-39	16.12
11	PG	1.02	0	194	-45	13.23
12	BA	5.46	0	228	-56	19.10
13	INTC	2.28	0	129	-81	12.57
14	TRV	-0.91	0	28	-43	3.82
15	CAT	0.50	0	130	-80	11.46
16	IBM	5.91	0	117	-65	13.98
17	UTX	3.18	0	123	-62	13.62
18	CVX	1.06	0	116	-134	10.12
19	JNJ	3.12	0	181	-87	15.27
20	VZ	3.33	0	196	-52	14.37
21	CSCO	2.01	0	173	-52	13.90
22	JPM	2.84	2	142	-62	15.67
23	WMT	0.88	0	126	-51	12.22

	Symbol	Mean	Median	Max	Min	Std. dev.
24	KO	2.34	0	179	-51	13.56
25	DIS	1.68	0	109	-65	10.42
26	UNH	1.46	0	149	-43	12.48
27	GS	1.80	1	236	-121	15.99
28	NKE	1.08	0	162	-196	15.90
29	V	1.87	0	182	-57	14.34

This table presents a summary of the descriptive statistics for the daily total news sentiment (Jan 2014–Jun 2016) for the 29 stocks used in this study. The summary statistics include mean value (Mean), median value (Median), maximum (Max), minimum (Min) and standard deviation (Std. dev.).

In Table 3.2 and 3.3, we can see that almost all the medians of the daily news volume for these 29 stocks are 0, as are almost all the medians of the daily news sentiment. It is difficult for us to assign weight to a portfolio based on this daily data. Therefore, we have decided to build the portfolio in response to the weekly data.

3.4 News items and portfolio selection

Portfolio performance measures are a key aspect of the investment decision–making process. Based on the idea of risk and return, a variety of evaluation techniques, such as the Sharpe ratio (Sharpe, 1966), the Treynor ratio (Treynor, 1965) and the alpha of Jensen (Jensen, 1969), were proposed and applied for evaluating the performance of the portfolio.

The Sharpe ratio is the most popular among them and this ratio has become the industry standard. It was developed by William F. Sharpe, the winner of the 1990 Nobel Memorial Prize in Economic Sciences. The Sharpe ratio is calculated as the difference between the mean portfolio return and the risk–free rate over the standard deviation of portfolio return.

$$\text{Sharpe ratio} = \frac{\bar{r_p} - r_f}{\sigma_p} \quad (3\text{-}1)$$

Here, $\bar{r_p}$ is the expected portfolio return, r_f is the risk–free rate and σ_p is the portfolio standard deviation. The Sharpe ratio is a risk–adjusted measure of return and it can be used to evaluate the performance of a portfolio. In this research, we use return, standard deviation and the Sharpe ratio to evaluate portfolios that have been built according to different methods.

For the risk free rate, there is no precise or widely accepted guidance on the appropriate debt maturity to use in modelling shareholder returns and risk premiums. Some researchers (Chawla, 2014, Brotherson et al., 2015) recommend selecting the yield to maturity on a long–term US government bond as a base interest rate. For this research, we use 2.5 years (Jan 2014–Jun 2016) as our data period; we then use the mean of the daily treasury yield curve rates for the 30–year government bonds from January 2014 to June 2016—which is 3.0017—as a proxy for the risk–free interest rate (our data source is the website of the US Department of the Treasury: *www.treasury.gov*).

To easily compare the results, we have annualised the return and the standard deviation. The annualised return formula is:

$$\text{Annualised Return} = ((\text{principal} + \text{gain}) \,/\, \text{principal})\ \text{^}\ (365/\text{days}) - 1 \quad (3\text{–}2)$$

To annualise the standard deviation, we simply multiply our daily standard deviation by the square root of the number of trading days.

$$\text{Annualised Standard Deviation} = \text{Standard Deviation of Daily Returns} * \text{Square Root (trading days)} \quad (3\text{–}3)$$

3.4.1 Can news volume contribute to portfolio selection?

Table 3.4 shows the weekly statistics for our 29 stock returns and Table 3.5 shows the weekly statistics for the news volume of our 29 stocks.

Table 3.4 The basic statistics for the weekly returns of the 29 stocks over 129 weeks from 8 January 2014 to 28 June 2016

	Symbol	Mean (%)	Median (%)	Max (%)	Min (%)	Std. dev. (%)
0	DJIA	0.059	0.305	5.597	–0.535	1.938
1	MMM	0.236	0.174	7.188	–6.549	2.208
2	DD	0.160	0.283	16.232	–9.915	3.599
3	MCD	0.251	0.184	7.512	–9.478	2.326
4	XOM	0.019	0.212	9.466	–11.797	2.828
5	MRK	0.175	0.305	6.303	–14.231	2.540
6	AXP	–0.259	0.269	7.422	–12.597	3.160
7	GE	0.172	0.195	11.070	–10.740	2.668
8	MSFT	0.354	0.299	15.291	–14.385	3.578
9	PFE	0.186	0.171	7.593	–11.768	2.459
10	HD	0.425	0.607	6.903	–9.634	2.706
11	PG	0.091	0.253	4.776	–8.931	1.942
12	BA	0.011	0.382	11.217	–13.216	3.384
13	INTC	0.265	0.488	9.713	–10.515	3.182
14	TRV	0.258	0.403	5.826	–9.340	2.178
15	CAT	–0.035	0.356	10.123	–10.282	3.376
16	IBM	–0.105	0.142	7.153	–11.192	2.841
17	UTX	–0.029	0.060	5.194	–10.673	2.495
18	CVX	–0.009	0.050	14.085	–16.083	3.828
19	JNJ	0.252	0.408	7.331	–7.996	2.084
20	VZ	0.195	0.476	8.622	–8.344	2.204
21	CSCO	0.279	0.389	16.821	–12.850	3.292
22	JPM	0.108	0.520	6.931	–12.168	2.993
23	WMT	0.011	0.236	7.896	–11.959	2.496
24	KO	0.151	0.384	6.388	–8.015	2.070
25	DIS	0.240	0.338	9.486	–11.250	2.916
26	UNH	0.536	0.429	9.816	–12.367	2.894

	Symbol	Mean (%)	Median (%)	Max (%)	Min (%)	Std. dev. (%)
27	GS	–0.103	0.297	7.185	–11.413	2.960
28	NKE	0.314	0.425	11.333	–10.983	3.088
29	V	0.294	0.591	12.118	–10.071	2.886

This table presents a summary of the descriptive statistics of the weekly return (from Wednesday to the following Tuesday from 8 Jan 2014–28 Jun 2016) for the 29 stocks used in this study. The summary statistics include the mean value (Mean), median value (Median), maximum (Max), minimum (Min) and standard deviation (Std. Dev.).

Table 3.5 The basic statistics for the weekly news volume for the 29 stocks over 129 weeks from 8 January 2014 to 28 June 2016

		News volume (all)					News volume (only new news)				
	Symbol	Mean	Median	Max	Min	Std. dev.	Mean	Median	Max	Min	Std. dev.
1	MMM	39.08	28	171	11	31.32	2.38	1	15	0	3.46
2	DD	66.01	42	493	8	66.92	2.39	1	15	0	3.59
3	MCD	126.20	104	450	30	75.64	2.93	1	20	0	4.13
4	XOM	145.14	130	416	53	72.55	2.70	2	17	0	3.78
5	MRK	87.83	77	331	16	53.16	3.40	2	21	0	3.89
6	AXP	99.47	86	319	25	53.14	2.68	2	14	0	3.15
7	GE	260.84	233	832	105	108.98	6.65	6	25	0	4.97
8	MSFT	299.97	277	850	131	106.29	4.79	4	20	0	4.20
9	PFE	151.48	117	617	42	103.70	3.75	2	19	0	4.19
10	HD	82.64	60	350	19	63.61	2.30	1	17	0	3.69
11	PG	99.33	85	340	30	58.92	3.09	2	20	0	3.81
12	BA	216.95	193	516	88	88.24	4.63	3	26	0	4.85
13	INTC	107.68	88	349	38	61.38	4.31	3	21	0	4.16
14	TRV	34.45	30	81	16	14.84	1.60	1	8	0	1.61
15	CAT	62.08	42	343	9	58.78	2.81	1	22	0	4.47
16	IBM	155.27	135	577	54	80.54	5.05	4	18	0	3.65
17	UTX	69.74	53	327	16	50.92	3.53	2	21	0	4.49
18	CVX	116.02	92	431	36	69.62	2.46	1	21	0	3.97
19	JNJ	111.46	102	364	31	57.57	3.14	2	21	0	3.85
20	VZ	167.53	155	384	42	69.48	3.75	3	17	0	3.73
21	CSCO	108.45	97	306	31	51.80	3.43	2	19	0	3.72
22	JPM	919.09	812	2421	263	409.50	20.85	19	87	1	12.69
23	WMT	193.55	164	680	59	97.57	4.16	3	24	0	4.71
24	KO	140.63	130	405	42	66.05	3.37	2	19	0	3.63

		News volume (all)					News volume (only new news)				
	Symbol	**Mean**	**Median**	**Max**	**Min**	**Std. dev.**	**Mean**	**Median**	**Max**	**Min**	**Std. dev.**
25	DIS	142.64	127	347	41	63.08	2.76	2	14	0	2.94
26	UNH	54.55	40	256	8	46.74	2.26	1	18	0	3.11
27	GS	748.20	628	2565	205	445.10	14.98	14	43	0	8.07
28	NKE	78.74	67	348	18	54.89	2.53	2	15	0	3.32
29	V	42.01	30	276	2	42.68	2.23	1	24	0	3.89

This table presents the summary descriptive statistics of the weekly news (all news and only new news) volume from Wednesday to the following Tuesday from 8 January 2014 to 28 June 2016 for the 29 stocks used in this study. The summary statistics include mean value (Mean), median value (Median), maximum (Max), minimum (Min) and standard deviation (Std. dev.).

The problem of portfolio selection can be considered as dealing with the situation in which an investor must determine how many shares of which assets to hold at which time instants in order to maximize the expected total utility from all consumption over the entire investment horizon (Korn and Korn, 2001). In other words, an investor must determine the weights of his or her portfolio components. In this section, we analyse the performance of the news volume–based portfolio selection strategy, following the power–law rule proposed by Kristoufek (2013) to obtain the weights of our portfolio components.

We use both in– and out–of–sample methods to assess the performance of the proposed methodology. The former is a standard approach for measuring the quality of portfolio optimisation; however, the latter is more useful for evaluating the practical application of the portfolio selection method. In this section (Section 3.4), the in–sample comprises portfolio weights that are built using information from the same period, while the out–of–sample consists of portfolio weights at Week t that are built using the information gleaned from Week t–1.

For each of the 29 stocks, let $V_{i,t}$ be the news volume for the stock–related term of stock i at Week t. The in–sample weight $w_{i,t}^{in}$ of stock i in the portfolio at Week t is defined as:

$$w_{i,t}^{in} = \frac{V_{i,t}^{-\alpha}}{\sum_{k=1}^{N} V_{k,t}^{-\alpha}} \quad (3–4)$$

Here, N is the number of stocks in the portfolio and α is a power–law parameter measuring the strength of discrimination for the stock volume. The normalisation factor $\sum_{k=1}^{N} V_{k,t}^{-\alpha}$ ensures that $\sum_{i=1}^{N} w_{i,t}^{in} = 1$for all t. From this definition, when $\alpha > 0$, stocks with more news are assigned a lower weight and where $\alpha < 0$, higher weights are attributed to stocks with more news. For $\alpha = 0$, a uniformly diversified portfolio is created where $w_{i,t}^{in} = \frac{1}{N}$.

For the out–of–sample, portfolio weights at Week t are built using the information gleaned from Week t–1. The out–of–sample weight $w_{i,t}^{out}$ of stock i in the portfolio at Time t is defined as:

$$w_{i,t}^{out} = \frac{V_{i,t-1}^{-\alpha}}{\sum_{k=1}^{N} V_{k,t-1}^{-\alpha}} \quad (3–5)$$

Here, N is the number of stocks in the portfolio and α is a power–law parameter measuring the strength of discrimination for the stock volume.

Figures 3.2 and 3.3 show the returns, standard deviations and Sharpe ratios for portfolios based on in–sample and out–of–sample portfolio performance judged according to the news volume (all news) approach for α and varying between –4 and 4 with a step of 0.1 respectively. The behaviours of the return, standard deviations and Sharpe ratios are practically identical for the in–sample and the out–of–sample: The returns rise with α; the standard

deviations decrease when α increases between $\alpha = -4$ and $\alpha = 0.6$ for the in–sample or between $\alpha = -4$ and $\alpha = 0.3$ for the out–of–sample where the deviation reaches its minimum; the Sharpe ratios likewise increase with α.

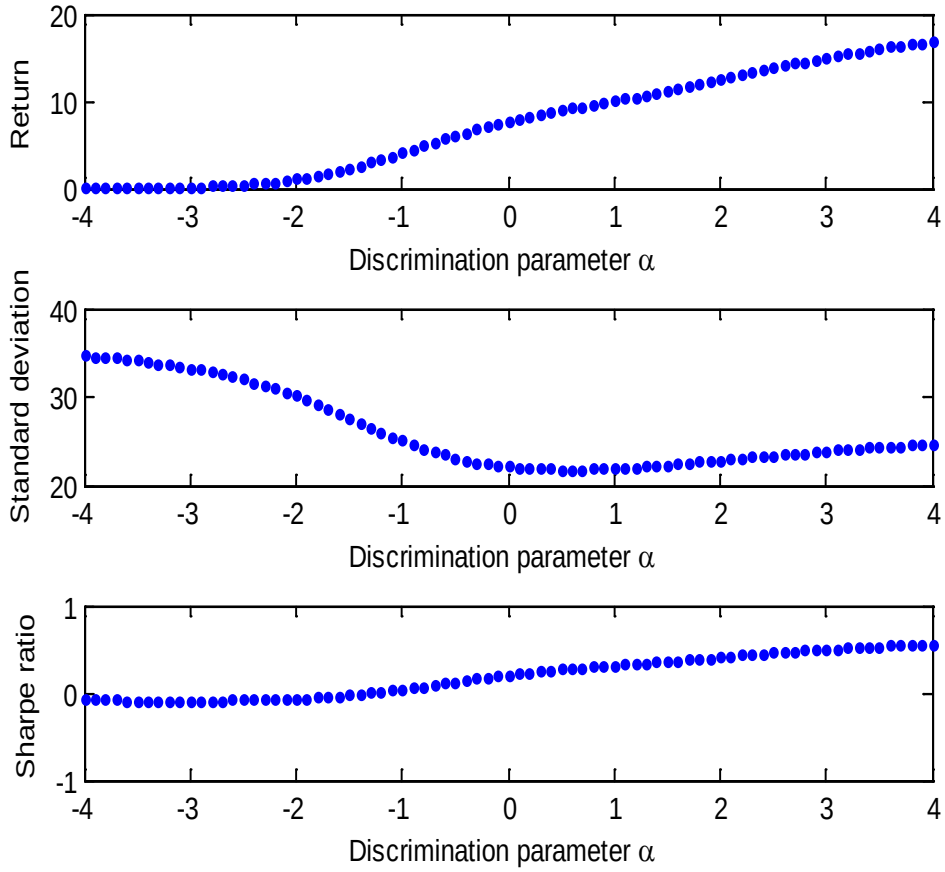

Figure 3.2 The in–sample portfolio performance based on news volume (all news)

Return, standard deviation and Sharpe ratio are shown for the in–sample performances of the constructed portfolio; the discrimination parameter α ranges between –4 and 4 with a step of 0.1. The middle point (α = 0) represents the uniformly weighted portfolio. The maximum return portfolio is found to be α = 4, which is the maximum value for α and the maximum return is 16.85%. The minimum return portfolio is found to be α = –0.19% and the minimum return is –3.30%. The maximum standard deviation (34.73%) portfolio is found to be α = –4, which is the minimum value for α. The minimum standard deviation portfolio is found to be α = 0.6 and the minimum standard deviation is 21.86%. The maximum Sharpe ratio portfolio is found to be α = 4 and the maximum value is 0.56. The minimum Sharpe ratio portfolio is found to be α = –3.1, while the minimum value is –0.08.

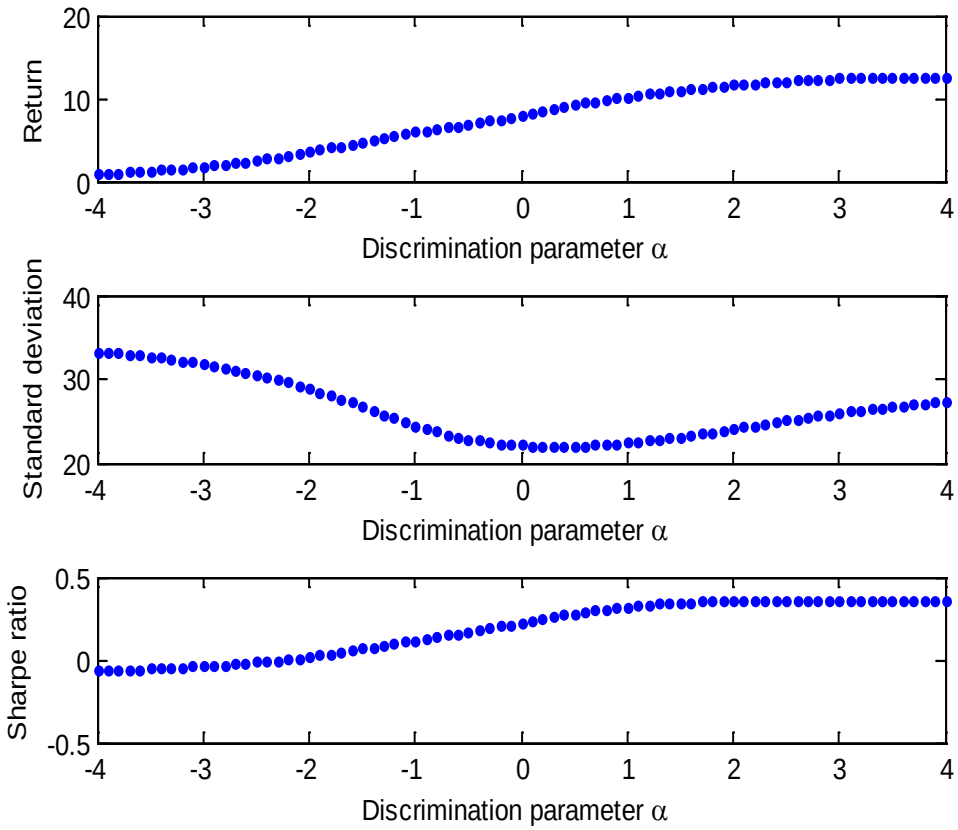

Figure 3.3 The out–of–sample portfolio performance based on news volume (all news)

Return, standard deviation and Sharpe ratio are shown for the out–of–sample performances of the constructed portfolio; the discrimination parameter α ranges between –4 and 4 with a step of 0.1. The middle point (α = 0) represents the uniformly weighted portfolio. The maximum return portfolio is found to be α = 3.9, while the maximum return is 12.67%. The minimum return portfolio is found to be α = –4 and the minimum return is 0.87%. The maximum standard deviation (33.30%) portfolio is found to be α = –4, which is the minimum value for α. The minimum standard deviation portfolio is found to be α = 0.3, while the minimum return is 22.02%. The maximum Sharpe ratio portfolio is found to be α = 2.8 and the maximum value is 0.36. The minimum Sharpe ratio portfolio is found to be α = –4 and the minimum value is –0.06.

Figures 3.4 and 3.5 show returns, standard deviations and Sharpe ratios for in–sample and out–of–sample portfolio performance based on the news volume (only new news) approach for α varying between –4 and 4 respectively with a step of 0.1. The behaviour of the standard deviations is

practically identical for the in–sample and the out–of–sample—the standard deviations decrease as α increases. For the in–sample, as α increases, the return decreases; the minimum return portfolio is found to be α = –1.4 and the minimum return is 4.62%. The return increases when $\alpha > 1.4$, while the maximum return portfolio is found to be α = 0.7 and the maximum return is 8.42%. After this, the return decreases again. For the out–of–sample, as α increases, the return likewise rises, yet the maximum return is 8.13% (α = 4) and the minimum return is –4.69% (α = –4). For the in–sample, the Sharpe ratio follows the changing of the return. As α increases, the Sharpe ratio decreases and the minimum Sharpe ratio is found to be α = –1.4, while the minimum value is 0.06. The Sharpe ratio increases when $\alpha > 1.4$, while the maximum Sharpe ratio portfolio is found to be α = 0.7 and the maximum value is 0.25. After this, the Sharpe ratio decreases again. For the out–of–sample, as α rises, the Sharpe ratio also increases. The maximum Sharpe ratio portfolio is found to be α = 4 and the maximum value is 0.23. The minimum Sharpe ratio portfolio is found to be α = –4, while the minimum value is –0.25.

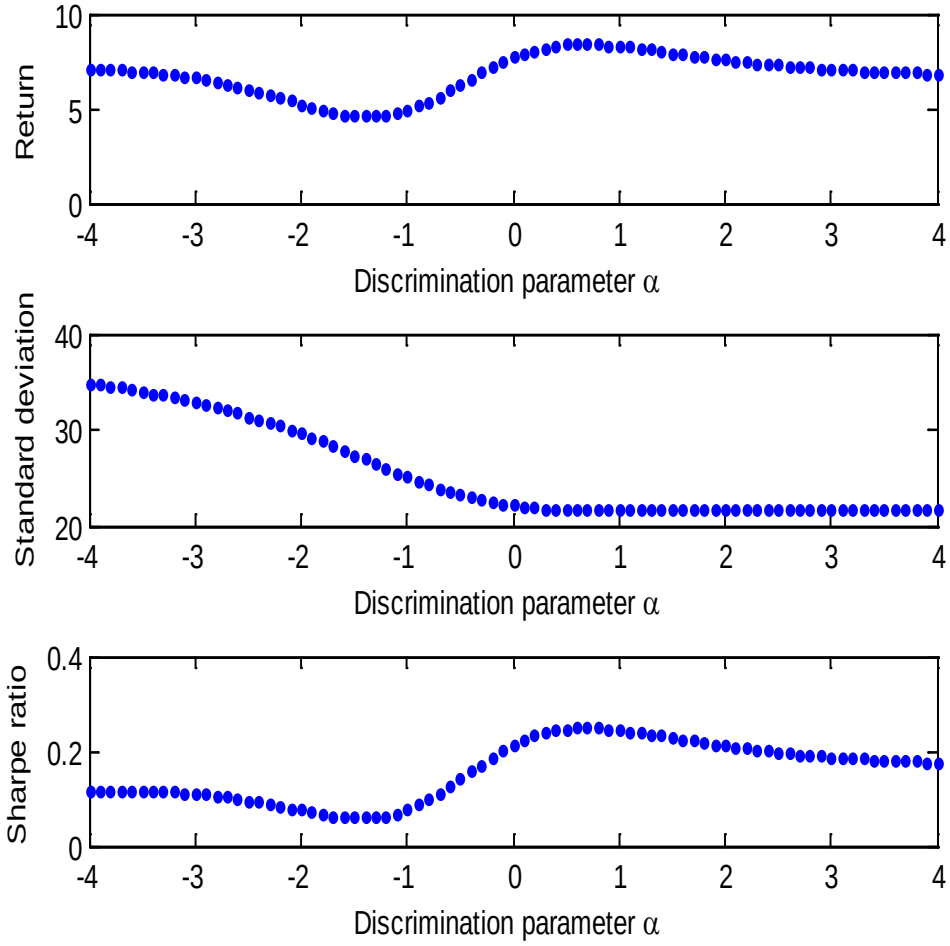

Figure 3.4 The in–sample portfolio performance based on the news volume (only new news)

Return, standard deviation and Sharpe ratio are shown for in–sample performances of the constructed portfolio; the discrimination parameter α ranges between –4 and 4 with a step of 0.1. The middle point (α = 0) represents the uniformly weighted portfolio. The maximum return portfolio is found to be α = 0.7 while the maximum return is 8.42%. The minimum return portfolio is found to be α = –1.4 and the minimum return is 4.62%. The maximum standard deviation (34.90) portfolio is found to be α = –4, which is the minimum value for α. The minimum standard deviation portfolio is found to be α = 1.1 and the minimum return is 21.61%. The maximum Sharpe ratio portfolio is found to be α = 0.7 and the maximum value is 0.25. The minimum Sharpe ratio portfolio is found to be α = –1.4 and the minimum value is 0.06.

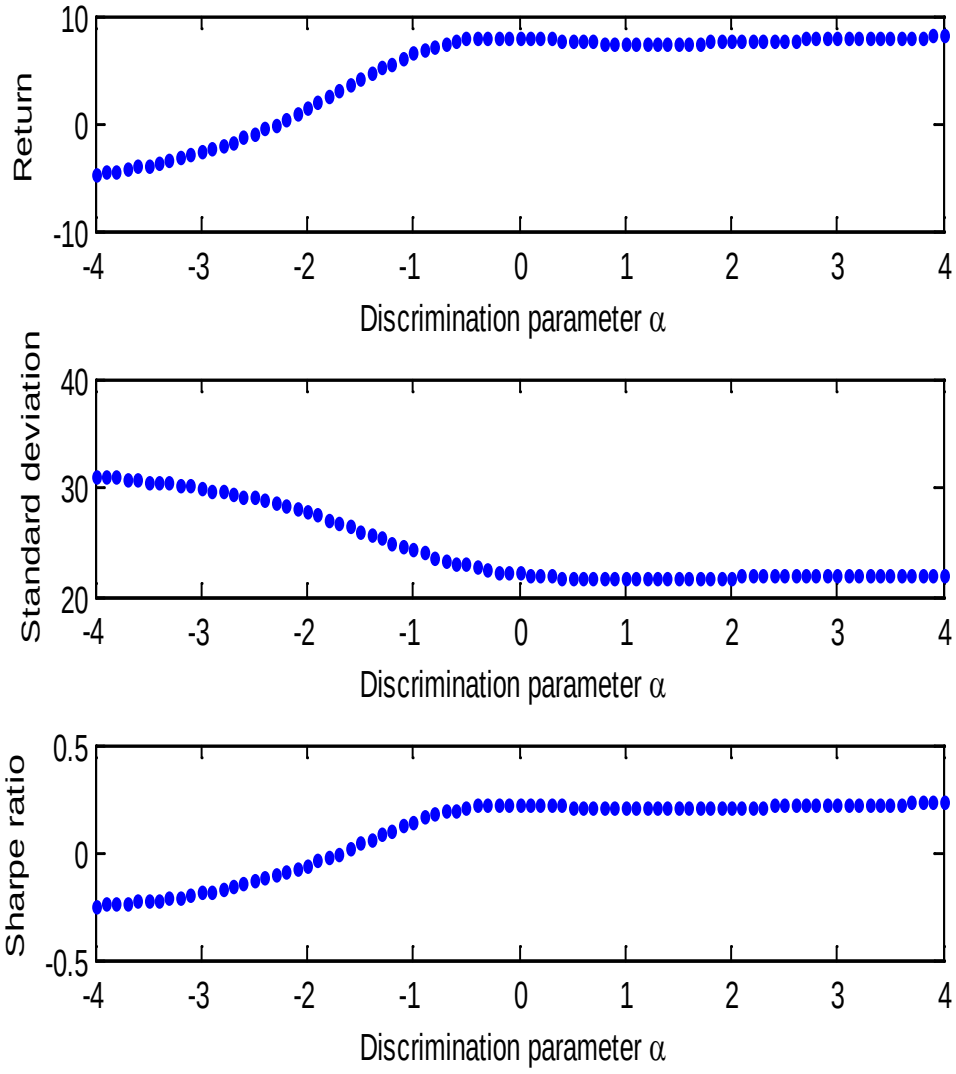

Figure 3.5 The out–of–sample portfolio performance based on the news volume (only new news)

Return, standard deviation and Sharpe ratio are shown for the out–of–sample performances of the constructed portfolio; the discrimination parameter α ranges between –4 and 4 with a step of 0.1. The middle point (α = 0) represents the uniformly weighted portfolio. The maximum return portfolio is found to be α = 4, while the maximum return is 8.13%. The minimum return portfolio is found to be α = –4 and the minimum return is –4.69%. The maximum standard deviation (31.15%) portfolio is found to be α = –4, which is the minimum value for α. The minimum standard deviation portfolio is found to be α = 1 and the minimum return is 21.75%. The maximum Sharpe ratio portfolio is found to be α = 4, while the maximum value is 0.23. The minimum Sharpe ratio portfolio is found to be α = –4 and the minimum value is –0.25.

3.4.2 Can news sentiment contribute to portfolio selection?

In this section, we consider the contribution of news sentiment when building portfolios, where $\mathrm{TS}_{\mathrm{i,t}}$ is the total news sentiment of stock i at Week t. $\mathrm{TS}_{\mathrm{i,t}} = \sum_{all\ news\ about\ stock\ i\ at\ week\ t} (\mathrm{ENS} - 50)$. Table 3.6 shows the weekly statistics for total news sentiment for the 29 stocks.

Table 3.6 The basic statistics for the weekly total news sentiment for the 29 stocks over 129 weeks from 8 January 2014 to 28 June 2016

	Symbol	**Mean**	**Median**	**Max**	**Min**	**Std dev**
1	MMM	10.23	0	126	–39	27.41
2	DD	8.21	0	179	–56	30.42
3	MCD	2.48	0	221	–78	33.42
4	XOM	6.30	0	79	–60	23.32
5	MRK	19.29	12	173	–39	32.77
6	AXP	9.98	2	131	–62	24.48
7	GE	67.28	53	364	–15	65.23
8	MSFT	27.50	22	139	–51	34.81
9	PFE	23.16	11	180	–60	39.83
10	HD	13.20	0	194	–38	41.13
11	PG	7.21	0	194	–45	35.64
12	BA	38.41	23	298	–65	53.72
13	INTC	15.69	8	176	–90	32.63
14	TRV	–6.46	0	18	–46	10.04
15	CAT	3.35	0	162	–115	30.50
16	IBM	41.67	36	169	–37	37.71
17	UTX	22.40	15	204	–61	36.84
18	CVX	7.52	1	116	–93	24.52
19	JNJ	21.87	12	181	–63	40.19
20	VZ	23.37	13	224	–52	42.25
21	CSCO	14.20	2	225	–57	39.50
22	JPM	19.85	9	194	–67	45.20
23	WMT	6.07	1	153	–51	32.56
24	KO	16.54	7	194	–55	35.16
25	DIS	11.87	4	104	–67	25.61

	Symbol	Mean	Median	Max	Min	Std dev
26	UNH	10.06	0	184	–53	32.19
27	GS	12.45	8	190	–121	40.09
28	NKE	9.02	0	209	–101	41.25
29	V	13.37	0	246	–57	44.40

This table presents the summary descriptive statistics of the weekly total news sentiment (from Wednesday to the following Tuesday over the period from 8 Jan 2014–28 Jun 2016) for the 29 stocks used in this study. The summary statistics include mean value (Mean), median value (Median), maximum (Max), minimum (Min) and standard deviation (Std. dev.).

We consider news sentiment according to the sorted order of total sentiment, where $\mathrm{SO_{i,t}}$ is the sorted order of $\mathrm{TS_{i,t}}$—the smallest $\mathrm{TS_{i,t}}$ with a value of 1, the largest $\mathrm{TS_{i,t}}$ with a value of 29. The in–sample weight $w_{i,t}^{in}$ of stock i in the portfolio at Time t is defined as:

$$w_{i,t}^{in} = \frac{\mathrm{SO_{i,t}^{-\alpha}}}{\sum_{\mathrm{k=1}}^{\mathrm{N}} \mathrm{SO_{k,t}^{-\alpha}}} \qquad (3\text{–}6)$$

Here, N is the number of stocks in the portfolio and α is a power–law parameter measuring the strength of discrimination for the stock sentiment. The normalisation factor $\sum_{\mathrm{k=1}}^{\mathrm{N}} \mathrm{SO_{k,t}^{-\alpha}}$ ensures that $\sum_{\mathrm{i=1}}^{\mathrm{N}} w_{i,t}^{in} = 1$ for all t. From this definition, when $\alpha > 0$, stocks with higher news sentiment are assigned a lower weight, but where $\alpha < 0$, we allocate heavier weights for stocks with higher news sentiment. For $\alpha = 0$, a uniformly diversified portfolio is desired, where $w_{i,t}^{in} = \frac{1}{\mathrm{N}}$.

The out–of–sample weight $w_{i,t}^{out}$ of stock i in the portfolio at Time t is defined as:

$$w_{i,t}^{out} = \frac{\mathrm{SO_{i,t-1}^{-\alpha}}}{\sum_{\mathrm{k=1}}^{\mathrm{N}} \mathrm{SO_{k,t-1}^{-\alpha}}} \qquad (3\text{–}7)$$

Figures 3.6 and 3.7 depict returns, standard deviations and Sharpe ratios for in–sample and out–of–sample portfolio performance based on the news

weekly sentiment approach for α and varying between –4 and 4 with a step of 0.1 respectively. The behaviour of the return is practically different for the in–sample and the out–of–sample; for the in–sample, as α rises, the return decreases; for the out–of–sample, when α increases, the return falls. The behaviour of the standard deviations is practically similar for the in–sample and the out–of–sample: when we assign more weights to the stocks with higher weekly sentiment, the portfolio has a smaller standard deviation. The Sharpe ratio follows the changes in returns as they are affected by the standard deviation. For the in–sample, the maximum Sharpe ratio portfolio is found to be $\alpha = -4$, while the maximum value is 0.70. When α increases, the Sharpe ratio decreases; the minimum Sharpe ratio is found to be $\alpha = 2.6$ and the minimum value is –0.47. For the out–of–sample, the minimum Sharpe ratio portfolio is found to be $\alpha = -4$, while the minimum value is 0.1335. As α rises, the Sharpe ratio also increases. The maximum Sharpe ratio portfolio is found to be $\alpha = 1.8$, while the highest value is 0.47.

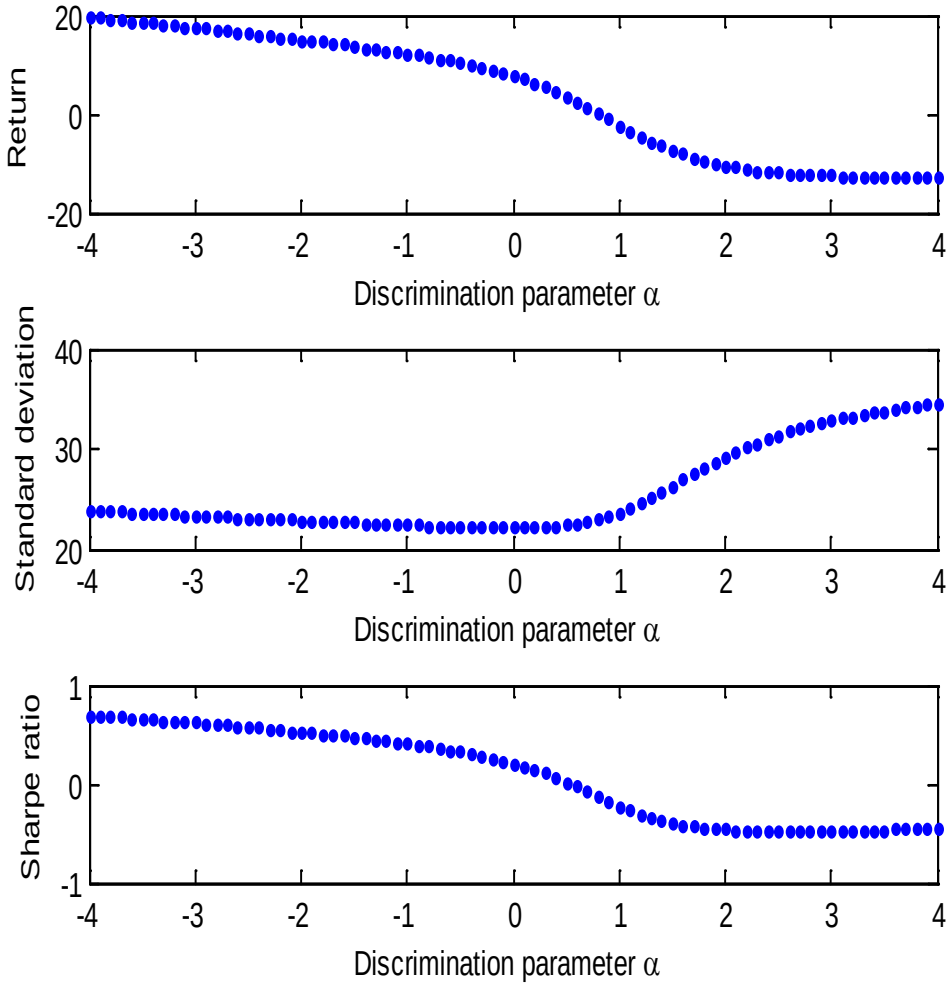

Figure 3.6 The in–sample portfolio performance based on the weekly news sentiment

Return, standard deviation and Sharpe ratio are shown for the in–sample performances of the constructed portfolio; the discrimination parameter α ranges between –4 and 4 with a step of 0.1. The middle point (α = 0) represents the uniformly weighted portfolio. The maximum return portfolio is found to be α = –4, which is the minimum value for α, while the maximum return is 19.68%. The minimum return portfolio is found to be α = 3.8, while the minimum return is –12.61%. The maximum standard deviation (34.51%) portfolio is found to be α = 3.8. The minimum standard deviation portfolio is found to be α = 0, while the standard deviation is 22.17%. The maximum Sharpe ratio portfolio is found to be α = –4, while the maximum value is 0.70. The minimum Sharpe ratio portfolio is found to be α = 2.6, while the minimum value is –0.47.

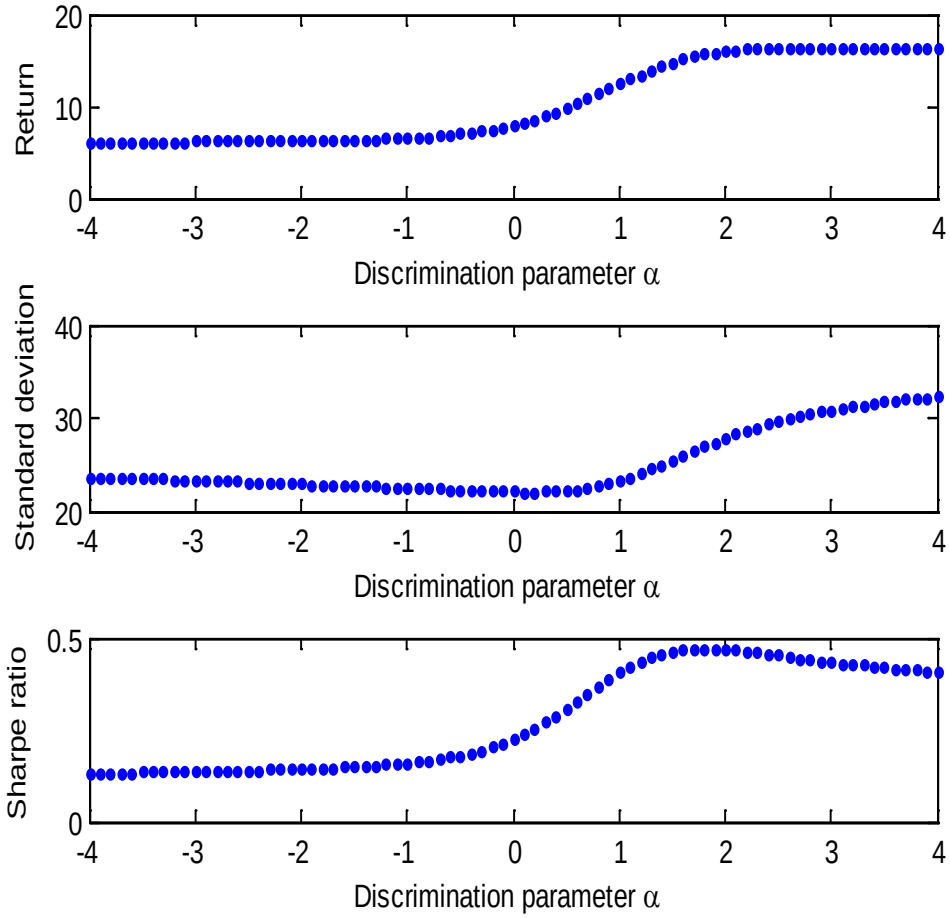

Figure 3.7 The out–of–sample portfolio performance based on the weekly news sentiment

Return, standard deviation and Sharpe ratio are shown for out–of–sample performances of the constructed portfolio; the discrimination parameter α ranges between –4 and 4 with a step of 0.1. The middle point (α = 0) represents the uniformly weighted portfolio. The maximum return portfolio is found to be α = 2.7, while the maximum return is 16.48%. The minimum return portfolio is found to be α = –4 and the minimum return is 6.17%. The maximum standard deviation (32.32%) portfolio is found to be α = –4, which is the minimum value for α. The minimum standard deviation portfolio is found to be α = 0.2, while the minimum return is 22.13%. The maximum Sharpe ratio portfolio is found to be α = 1.8, while the maximum value is 0.47. The minimum Sharpe ratio portfolio is found to be α = –4 and the minimum value is 0.1335.

3.4.3 Conclusions and discussion

Our research shows some interesting results concerning the relationship between news volume, news sentiment and portfolio performance. For news volume, the results show that if we assign greater weight to stock with higher

levels of news volume (all news and only new news), the portfolio has higher standard deviation. This means that high news volume contributes to a portfolio's risk. For news sentiment, positive news sentiment contributes to the portfolio return in–sample, while negative news sentiment contributes to the portfolio return out–of–sample, which occurs as a consequence of investors overreacting to the news sentiment.

Our results enhance the literature in two ways. First, we contribute to the discussion about the relationship between news volume and the stock market. There is little research to address the effect of the news volume on the stock market and the existing research only considers some types of news. For instance, Alanyali et al. (2013) exploit a large corpus of daily print issues of the *Financial Times* and find a positive correlation between the number of daily mentions of a particular company in the *Financial Times* and the daily transaction volume of that company's stock both on the day before the news is released and on the same day as the news is released. This research considers news relating to a single firm, whereas our research found that all news contributes to the portfolio return. This result is consistent with those of Gillam et al. (2015), who propose a measure of abnormal news volume that controls for the size of the firm and the analyst attention that it receives, demonstrating that news volume information can enhance returns.

Further, our research contributes to the exploration of the relationship between news sentiment and the stock return, which has been discussed by several studies. For instance, Heston and Sinha (2017) find that daily news can be used to predict stock returns; Allen et al. (2015) show that news sentiment score contains useful information about factors impacting on the

volatility of the DJIA; Ho and Wang (2016) develop an ANN model to predict the stock price movements of GOOG using the news sentiments for Google Inc.

Additionally, the different portfolio performances in and out–of–sample can be explained as the investors overreacting to the news sentiment, a phenomenon that has been discussed by other scholars. For instance, Barberis et al. (1998) find that stock prices are considered to have been altered by the overreaction of investors if the average return that follows not one but a series of announcements of good news is lower than the average return that follows a series of negative news stories. Boubaker et al. (2015) find evidence of short–term overreaction in the Egyptian stock exchange where losers ('bad news' portfolios) significantly outperform winners ('good news' portfolios).

3.5 A proposed new portfolio selection method based on the kNN

3.5.1 Theoretical background: The kNN for classification

The k–Nearest–Neighbor (kNN) is one of the most fundamental and simple classification methods based on the closest training examples in the feature space. An unknown pattern can be classified according to the majority vote of its neighbors. It is one of the first choices for a classification study as it needs little or no prior knowledge about the distribution of the data.

The kNN has been used in many applications such as face recognition (Yang, 2006, Masip and Vitrià, 2008), handwriting recognition (Kumar et al., 2011, Zanchettin et al., 2012), text classification (Han et al., 2001, Yong et al., 2009) and forest field plot (Haapanen and Ek, 2001, Reese et al., 2002).

The concept of kNN method is quite simple. The k is a positive integer,

typically small. Given a new unlabelled sample, the system finds the k nearest neighbors among the training samples. In other words, each sample is classified by a majority vote of its neighbors. For example, if k = 1, the sample is simply assigned to the class of its nearest neighbor. In a two–class classification problem, k is normally an odd number to avoid tied votes. Figure 3.8 shows the use of kNN where k = 3. In this example, the three nearest neighbors are a given unlabelled sample denoted as ☆ and two adjacent samples, denoted as △, which belong to Class 1. The given sample ☆ also has one adjacent sample, denoted as ○, from Class 2. Hence, by following the rule of majority vote, the unlabelled sample ☆ will be assigned to Class 1.

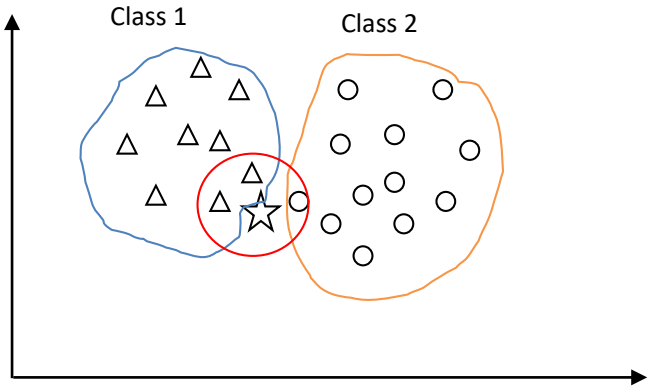

Figure 3.8 An illustration of sample classification using the kNN method when k=3.

To achieve accurate classification, the prime concern when using the kNN is how to define ‘nearest’; in other words: how to find a smart way to measure the similarity of samples. Researchers attribute different similarity functions to different problems. For example, in text classification, researchers use cosine distance (Tan, 2006, Yu and Yu, 2007):

$$\mathrm{Sim}(d_1, d_2) = (\sum_{l=1}^{V} d_{1l} * d_{2l}) / (\sqrt{\sum_{l=1}^{V} d_{1l}^2} \sqrt{\sum_{l=1}^{V} d_{2l}^2}) \qquad (3\text{–}8)$$

Here, V denotes the dimensions of a document with vectors d1 and d2; kNN training is extremely fast since it needs only to calculate the distance.

In these years, kNN method have been used in the research of finance. For instance, Teixeira and de Oliveira (2010) propose a method for automatic stock trading that combines technical analysis and the nearest neighbor classification.

3.5.2 A new portfolio selection method based on the kNN

The kNN classifier is a machine learning algorithm that is considered simple to implement. In this type of classifier, a new pattern is classified according to its similarity with the available training patterns. The performance of a kNN classifier is primarily determined by the choice of K as well as by the distance metric applied. The most crucial aspect of kNN is how to define ‘nearest’.

The data sets can be classified as the training data set and the test data set. A training set is a set of data used to discover potential relationships. A test set is a set of data used to assess the strength and utility of the proposed kNN method. To measure the similarity between the two, the Euclidean distance between the data in the test data set and that in the training data set is computed. Next, the class of the training pattern with the smaller distance is assigned to the test data.

The basic idea behind this method is “history repeats itself” and future market direction can be determined by examining past patterns while all

technical analysis rests on this assumption. In section 3.4, we conclude positive news sentiment contributes to the portfolio return in–sample, while negative news sentiment contributes to the portfolio return out–of–sample. In this section, we use news sentiment and stock returns to calculate the distance.

For this research, we consider the simplest situation and choose k = 1. Figure 3.9 shows the structure of the kNN portfolio selection method. At the beginning of every week in this research (i.e., the opening time of the stock market on Wednesdays), we calculate the distance of Stock i from the other stock in the training data set. We select the return with the smallest distance as the predicted return of Stock i for each week and sort the predicted returns of each of the 29 stocks, assigning the sorted order $\mathrm{SO}_{\mathrm{i,t}}$ of 29, 28 … 1 to them. The smallest expected return has the value 1, while the largest expected return has the value 29. The weight $w_{i,t}$ of Stock i in the portfolio at Time t is defined as:

$$w_{i,t} = \frac{\mathrm{SO}_{\mathrm{i,t-1}}^{-\alpha}}{\sum_{\mathrm{k}\,=\,1}^{\mathrm{N}} \mathrm{SO}_{\mathrm{k,t-1}}^{-\alpha}} \tag{3–9}$$

We update the portfolio weight every week during the test period using Equation 3–9.

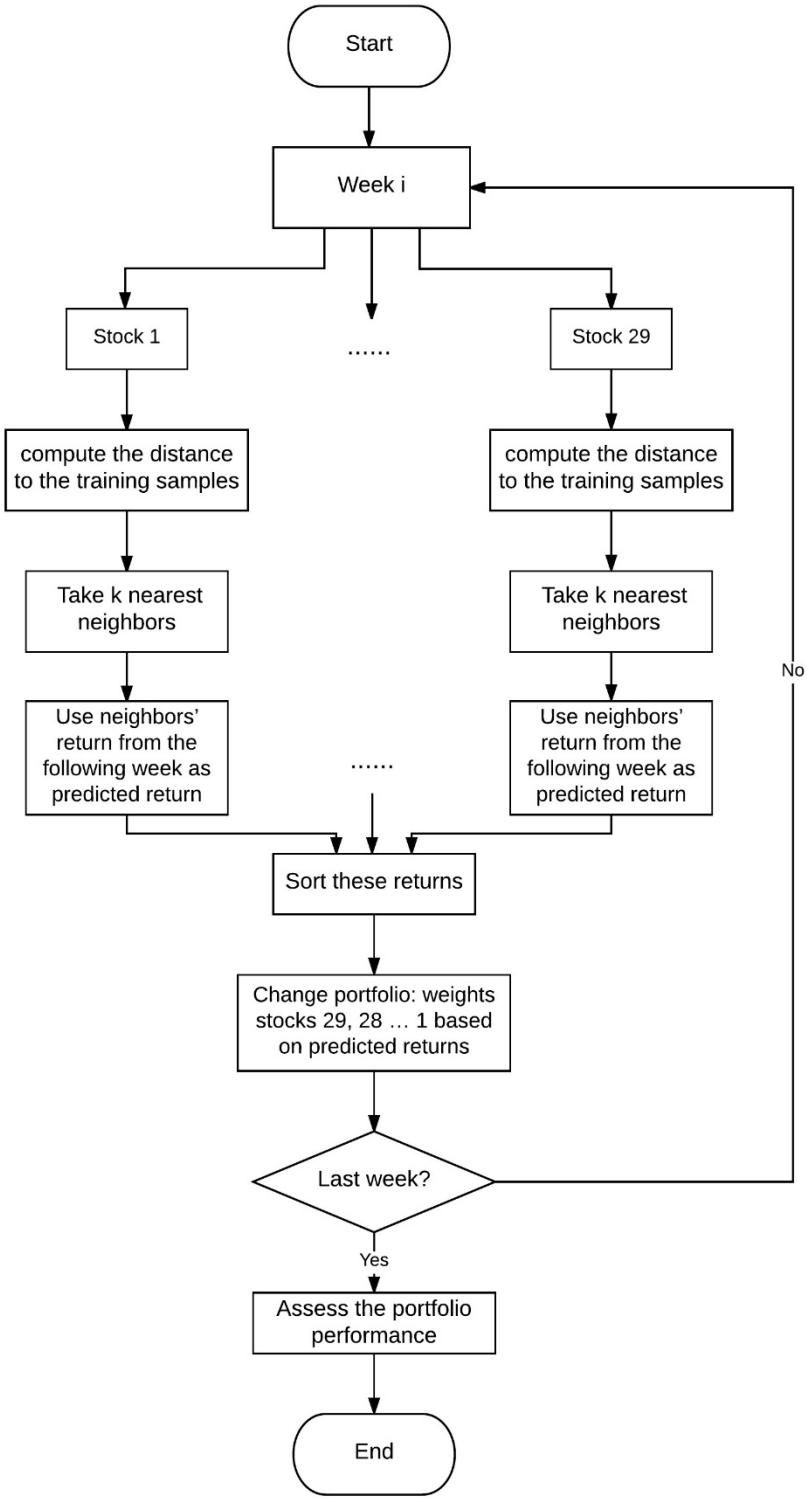

Figure 3.9 The structure of the kNN portfolio selection method

3.5.3 Empirical results

In Section 3.4, we define the in–sample as the attribution of portfolio weights using information from the same period and the out–of–sample as the allocation of portfolio weights at Week t using the information gleaned from Week t–1. In this section, we use different definitions for in–sample and out–of–sample. For this research, we consider only 129 weeks (about 2.5 years), from 8 January 2014 to 28 June 2016. The first 121 weeks (8 Jan 2014–5 Jan 2016) are used as the in–sample period (also called the training period) and the last 25 weeks (6 Jan 2016–25 Jun 2016) are used as the out–of–sample period (also called the test period).

We consider only 29 of the 30 components of the DJIA index; we do not include Apple Inc. (NASDAQ: AAPL) or AT&T Inc. (NYSE: T) in our research (Apple Inc. replaced AT&T Inc. on 19 March 2015). We use the out–of–sample period to evaluate the performance of the proposed method and we are mainly interested in three portfolio performance measures: return, standard deviation and the Sharpe ratio. Standard deviation is a common measure of risk and the Sharpe ratio represents the standardised average return of the portfolio.

Before we assess the performance of the proposed kNN method, we will discuss the out–of–sample performance of the mean–variance method. Figure 3.10 shows the frontier of the mean–variance method during the in–sample period, which reveals the balance between the return and the standard deviation (risk). Table 3.7 shows the out–of–sample performance of the mean–variance method. For instance, for the portfolio with an in–sample annualised return of 22% and the minimum standard deviation, we retain the weights of the portfolio and assess its performance during the out–of–sample period. The

out–of–sample annualised return, the annualised return standard deviation and the Sharpe ratio are –0.38%,14.40% and –0.23 respectively.

Our results indicate poor out–of–sample performances for the mean–variance portfolio optimisation method. The results are consistent with the findings of other scholars, (e.g., Bera and Park (2008) DeMiguel et al. (2009)). Further, our results show that using Markowitz's mean–variance portfolio optimisation method leads to portfolios that are highly concentrated on a few assets. We consider 29 stocks to build the portfolio and we can see for the in–sample return of 4% that the portfolio with the minimum standard deviation is built with only 12 stocks. This result concurs with the findings of Bera and Park (2008).

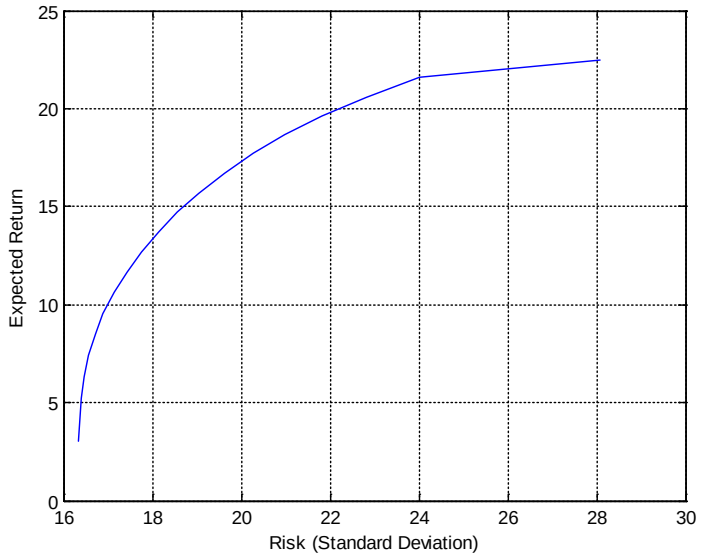

Figure 3.10 The frontier of the mean–variance method during the training period

Table 3.7 The out–of–sample performance of the mean–variance method

In–sample			**Out–of–sample**			**Stock numbers**
Return (%)	**Std. (%)**	**Sharpe ratio**	**Return (%)**	**Std. (%)**	**Sharpe ratio**	
4	16.35	0.43	23.58	8.22	2.50	12
6	16.44	0.55	22.81	8.28	2.39	11
8	16.64	0.66	20.99	8.47	2.12	11
10	16.99	0.77	18.34	8.78	1.75	9
12	17.53	0.86	15.96	9.23	1.40	10
14	18.26	0.93	12.33	9.84	0.95	8
16	19.21	0.99	8.73	10.57	0.54	7
18	20.48	1.03	6.75	11.35	0.33	7
20	22.18	1.04	4.68	12.45	0.14	6
22	24.84	1.01	–0.38	14.40	–0.23	4

As discussed, the mean–variance portfolio optimisation method leads to poor performance during the test period. To assess the proposed kNN method, we use the DJIA index for the benchmark. Table 3.8 shows the out–of–sample performance of the DJIA index and the kNN portfolio selection method. The return for the kNN method is 19.17% while the return for the DJIA index portfolio is only 3.70%. The standard deviation of the kNN method is 9.93%, while the standard deviation of the DJIA index portfolio is 9.97%. The proposed kNN method dominates the DJIA index portfolio both in terms of return and standard deviation. To further illustrate this, Figure 3.11 compares the evolution of the kNN portfolio to that of the DJIA index. The value of the out–of–sample kNN portfolio at the end of the analysed period is 108.8% of its initial value, which corresponds to a cumulative profit of 8.8%. In contrast to the DJIA index, which has a cumulative profit of 1.8%, the kNN strategy yields approximately five times the profit.

Table 3.8 The performance of the kNN portfolio selection method

	DJIA index portfolio	**kNN portfolio**
Return (%)	3.70	19.17
Standard deviation (%)	9.97	9.93
Sharpe ratio	0.07	1.63

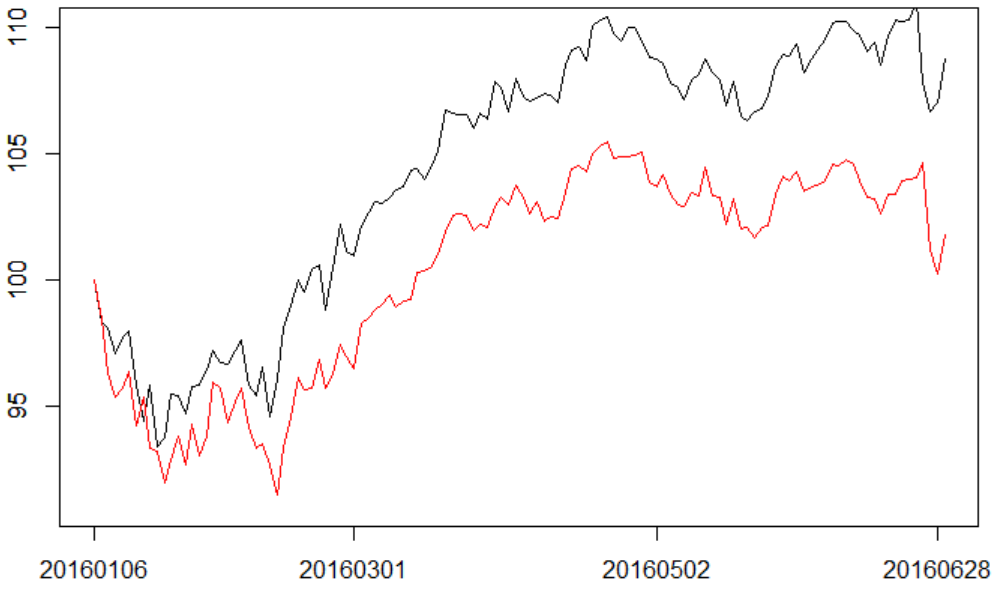

Figure 3.11 The evolution of portfolio value based on the kNN method using news sentiment

The red line represents the evolution of the DJIA index; the black line shows the performance of the out–of–sample diversification using the proposed kNN method. The opening price of 6 Jan, 2016 as the benchmark (value 100). Portfolio value is shown on the y–axis. The comparison of the black and red lines is essential as it shows the significant profits to be made if we apply the kNN–based strategy.

3.5.4 Robustness checks

To ascertain the robustness of the proposed method, we access the kNN portfolio performance for every week of the out–of–sample period. The results are shown in Table 3.9. Of these 25 weeks, there were 17 weeks (68%) where the return of the kNN portfolio was better than that of the DJIA index portfolio. The results show the robustness of the proposed method:

Table 3.9 The robustness of the kNN portfolio selection method

	Return		
Week	**DJIA index portfolio**	**kNN portfolio**	**Better than DJIA?**
1	–3.66	–2.06	YES
2	–4.59	–4.32	YES
3	2.54	2.17	NO
4	0.11	0.98	YES
5	–0.93	–0.12	YES
6	2.79	3.53	YES
7	–0.39	–1.15	NO
8	2.63	3.33	YES
9	0.70	1.41	YES
10	1.65	0.90	NO
11	1.97	1.99	YES
12	0.36	1.21	YES
13	–0.27	–0.72	NO
14	0.77	1.32	YES
15	1.79	1.67	NO
16	–0.35	–0.24	YES
17	–1.45	–2.04	NO
18	1.04	0.92	NO
19	–2.33	–2.08	YES
20	1.34	1.82	YES
21	0.30	0.28	NO
22	0.80	1.37	YES
23	–1.27	–0.78	YES
24	0.73	0.82	YES

	Return		
Week	**DJIA index portfolio**	**kNN portfolio**	**Better than DJIA?**
25	–2.11	–1.36	YES
			68%

3.6 Conclusions

The current, standard approach to portfolio selection based on the mean and variance of the assets often leads to a lopsided concentration on a few firms and poor out–of–sample forecasting performance. The starting point of this research was a curiosity towards the connection between strategic decision making and news. As a consequence, we propose two major modifications to the current approach. The first modification involves the use of big datasets, such as news volume and news sentiment scores associated with the firms. This modification is motivated by empirical evidence that news volume and sentiment can significantly affect asset return and risk. The other modification involves the application of the kNN algorithm, which is commonly used in the classification and regression of large datasets. The proposed kNN method is extremely fast, since it consists solely of the storage of all training patterns. This is a great advantage for our proposed method as we intend to analyse a high number of stocks on a daily basis.

Our results indicate that news volume and sentiment can enhance the current approach to portfolio selection. In particular, in–sample and out–of–sample tests suggest that the proposed kNN portfolio selection approach dominates the benchmark DJIA index portfolio.

Chapter 4:

Essay 3: Combined Markov and hidden Markov model in stock price movement prediction

4.1 Introduction

The literature has provided various explanations for the movement of stock prices. The most important of them is the random walk hypothesis and the efficient market hypothesis. The random walk hypothesis (Fama, 1965) states that stock market prices evolve according to a random walk. The efficient market hypothesis (Fama, 1970) claims that securities markets are extremely efficient in reflecting information about individual stocks and about the stock market as a whole. In other words, the stock price is determined by all relevant information.

Conversely, many researchers show that stock price fluctuations depend on other factors, including interest rate (Christie, 1982, Flannery and James, 1984, Alam and Uddin, 2009), insider information (Kyle, 1985, Wang and Wang, 2017), unexpected extreme news (Chan, 2003, Asgharian et al., 2011), prescheduled earnings announcements (Jennings and Starks, 1986, Skinner, 1994, Su, 2003), political events (Kim and Mei, 2001, Amihud and Wohl, 2004, Jensen and Schmith, 2005) and corporate takeovers (Malatesta and Thompson, 1985, Franks and Harris, 1989, Pound and Zeckhauser, 1990).

Quite a long time ago, speculators, investors and traders use technical analysis to try to predict the stock price movement (Abu-Mostafa and Atiya,

1996). The practice of technical analysis evaluates securities by analysing the statistics generated by market activity, such as past prices and volume. Speculators, investors and traders employ charts (Leigh et al., 2008), technical indicators (Tanaka-Yamawaki and Tokuoka, 2007), oscillators (Koutmos, 1996, Cohen and Cabiri, 2015) and other tools to identify patterns that can suggest future activity. The key assumption of the technical analysis is that 'history tends to repeat itself'.

The enormous amount of valuable data generated by the stock market has encouraged researchers to attempt prediction using different methodologies. With the development of the computer and computing techniques, there is a burgeoning strand of literature on the application of data mining techniques to the analysis of stock price movements (Atsalakis and Valavanis, 2009, Hajizadeh et al., 2010). These data mining techniques include decision trees (Wang and Chan, 2006, Wu et al., 2006, Chang, 2011), clustering (Harris, 1991, Lai et al., 2009), ANNs (Wong and Selvi, 1998, Paliwal and Kumar, 2009, Ho and Wang, 2016) and the support vector machine (Tay and Cao, 2001, Huang et al., 2005, Ni et al., 2011), etc.

There is an increasing body of research about the application of the Markov models and HMMs to finance. Markov models were first proposed by Andrei Markov who studied them in the early twentieth century. They are used to model randomly changing systems wherein future states depend only on the current state and not on events that have occurred previously. Based on the idea of Markov property, Hamilton (1989) proposed the regime–switching model, which involves multiple structures (equations) that can characterise time series behaviours in different regimes. The regime–switching model is

able to capture more complex dynamic patterns and is widely used in economics and finance.

HMMs describe the relationship between two stochastic processes: an observed process and an underlying, ‘hidden’ (unobserved) process. The hidden process is assumed to follow a Markov chain and the observed data are modelled as independent, yet conditional on the sequence of hidden states. After successful applications in speech and handwriting, HMMs have been employed in financial market prediction by some researchers (Hassan and Nath, 2005, Gupta, 2012, Lee et al., 2014). Nonetheless, most of these studies use stock return for the observations of the HMM models, which conflicts with the assumption that the observations are independent.

In this research, we propose a new model wherein the observation is affected by a Markov model and an HMM model. The proposed model better describes the nature of the stock market and exhibits better potential prediction ability. In the first section of this chapter, the structure of the proposed model is described; after this, problems in the evaluation, decoding and learning within the proposed model are discussed, along with how to solve them. Next, we explore potential applications of the proposed model in the stock market by designing trading strategies based on it.

The remainder of the chapter is organised as follows. In the second section, we introduce the research background. In Section 4.3, we describe the structure of the proposed model and discuss the problems in evaluation, decoding and learning. The application of the proposed model for stock market price prediction is discussed in Section 4.4. The final section concludes the chapter.

4.2 Research background

4.2.1 Markov model

In probability theory, a Markov model is a stochastic model that consists of a list of the possible states of a system, the possible transition paths between those states and the rate parameters of those transitions. In a Markov model, it is assumed that future states depend only on the current state, not on events that occurred before it (this is called the Markov property).

Markov models assume that there are a finite number of discrete states, which are called Markov states. If we suppose that the states are numbered and that $\mathrm{T} = \{1, 2, \ldots\ \mathrm{t}\}$ denotes the set of transient states, the transition probabilities can be described as:

$$\boldsymbol{P}_T = \begin{bmatrix} P_{11} & \ldots & P_{1t} \\ \ldots & \ldots & \ldots \\ P_{t1} & \ldots & P_{tt} \end{bmatrix} \tag{4–1}$$

The value $\mathrm{P_{ij}}$ represents the probability that the process will, when in state i, make a transition into state j. It is understandable that probabilities are nonnegative and that the process must make a transition into some state. We have that:

$$\mathrm{P_{ij}} \geq 0, \mathrm{i}, \mathrm{j} \geq 0; \sum_{\mathrm{j}=0}^{\mathrm{t}} \mathrm{P_{ij}} = 1\,, \mathrm{i} = 1,2..\mathrm{t} \tag{4–2}$$

Figure 4.1 shows an example of the use of a Markov model to describe a hypothetical stock market. The states represent whether the hypothetical

stock market is exhibiting uptrend, downtrend or sideways trend on a given day. According to the figure, an uptrend day is followed by another uptrend day 50% of the time, a downtrend day 10% of the time and a sideways trend day the other 40% of the time. Labelling the state–space (1 = downtrend, 2 = sideways, 3 = uptrend) the transition matrix for this example is:

$$\boldsymbol{P}_T = \begin{bmatrix} 0.5 & 0.4 & 0.1 \\ 0.3 & 0.4 & 0.3 \\ 0.1 & 0.4 & 0.5 \end{bmatrix} \qquad (4\text{–}3)$$

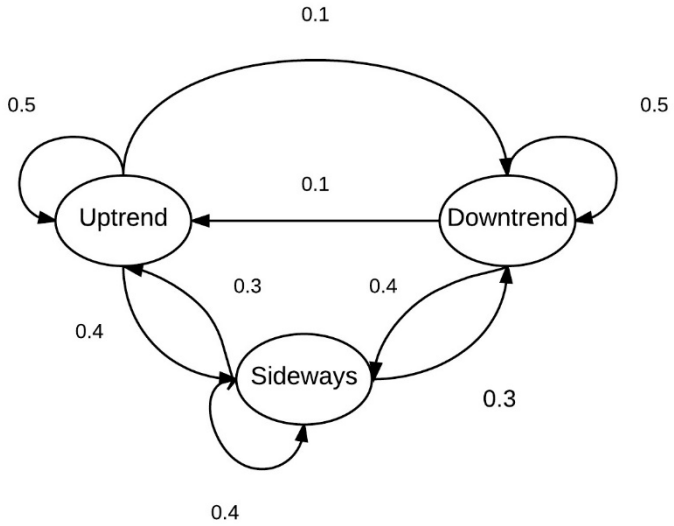

Figure 4.1 An example of the use of a Markov model to describe a hypothetical stock market
Each circle represents a Markov state. There are three states: uptrend, downtrend and sideways in the stock market. Arrows indicate allowed transitions.

Markov processes can be similarly classified according to the type of time parameter and the type of state–space parameter. For instance, a discrete–time and discrete–state Markov process is called a discrete–time Markov chain. A Markov chain is classified as non–homogeneous if the future state is

dependent on the time parameter. It is called homogeneous if it is independent of time.

4.2.2 Regime–switch model

In the Markov regime–switching model, the regime switching is affected by a state variable that follows a first–order Markov chain. This means that the current value of the state variable is immediately affected by the value of the last period. A two–state Markov–switching AR(1) model, which follows Hamilton (1989) structure, is shown as:

$$Z_t = \alpha_0 + \alpha_1 s_t + \beta Z_{t-1} + \varepsilon_t \quad (4\text{–}4)$$

$$\varepsilon_t \sim \text{i.i.d.}\, N(0, \sigma^2) \quad (4\text{–}5)$$

$$\Pr[S_t = 1 | S_{t-1} = 1] = p \quad (4\text{–}6)$$

$$\Pr[S_t = 0 | S_{t-1} = 0] = q \quad (4\text{–}7)$$

Following Hamilton's work, a number of extended Markov regime–switching models have been proposed, such as the continuous–time Markov regime–switching model (Zhou and Yin, 2003), the regime–switching model with time–varying parameters (Mount et al., 2006) and the regime–switching long memory model (Haldrup and Nielsen, 2006).

Regime–switching models have long been a tool available to economics and finance. Regime–switching models with constant transition probabilities have been applied to interest rates (Gray, 1996, Dahlquist and Gray, 2000, Ang and Bekaert, 2002), the behaviour of gross national product (Durland and McCurdy, 1994, Clements and Krolzig, 1998, Lam, 2004), option valuation (Bollen, 1998, Buffington and Elliott, 2002, Henriksen, 2011, Shen et al., 2014),

portfolio selection (Zhou and Yin, 2003, Elliott et al., 2010, Hua and Wang, 2014), speculative bubbles (Van Norden and Schaller, 1999, Al-Anaswah and Wilfling, 2011, Shi and Song, 2014) and foreign exchange rates (Engel, 1994, Bollen et al., 2000, Marsh, 2000).

There is a significant amount of research utilising regime–switching models in the financial market. Dueker (1997) applies switching conditional variance models to financial markets and examines their multi–period stock market volatility forecasts as predictions of options–implied volatilities. Alizadeh and Nomikos (2004) propose a new approach for determining time–varying minimum variance hedge ratios in stock index futures markets by using Markov regime–switching models. Moore and Wang (2007) investigate the volatility of stock markets in the new European Union member states by utilising the Markov regime–switching model. Timmermann (2012) develops an asset–pricing model that represents breaks in the context of a Markov–switching process with an expanding set of nonrecurring states. The model presents empirical evidence on the existence of structural breaks in the fundamental processes underlying US stock prices. Zhu and Zhu (2013) introduce a regime–switching combination approach to predict excess stock returns. They find that excess returns are more predictable during economic contractions than during expansions.

Considering news sentiment is a new direction in the financial market prediction. Chung et al. (2012) implement a multivariate Markov–switching model to capture the unobservable dynamics of the changes in the economic regime and examine asymmetries in the predictive power of investor sentiment about the cross–section of stock returns in economic expansion and recession

states. Ho et al. (2013) examine the dynamic relationship between firm–level return volatility and public news sentiment using two–state Markov regime–switching GARCH (generalized autoregressive Conditional heteroscedasticity) models. Their results show the significant effect of firm–specific news sentiment on intraday volatility persistence.

4.2.3 Hidden Markov Model

An HMM is a finite state machine with a fixed number of states that provides a framework for modelling a time series of multivariate observations that are probabilistic with internal states that are either hidden or not directly observable. HMMs were introduced at the beginning of the 1970s as a tool in speech recognition (Rabiner, 1989). This model, based on statistical methods, has become increasingly popular over the last several years as a consequence of its strong mathematical structure and theoretical basis for use in a wide range of applications.

Figure 4.2 shows the general structure of an HMM—a doubly stochastic process in which the underlying stochastic process is unobservable (S_i); in other words, the states are hidden. However, there is another stochastic process (based on the hidden states) that produces a sequence of observations (O_i):

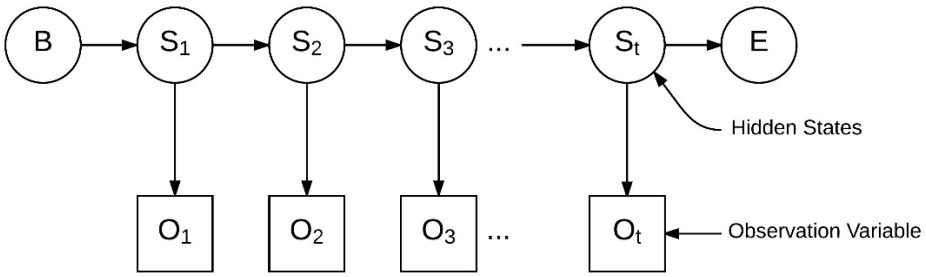

Figure 4.2 The general structure of an HMM

HMMs have been used in analysing and predicting time series phenomena and have been applied to many different areas, including speech recognition (Rabiner, 1989, Huang et al., 1990), gene profiling and recognition (Lukashin and Borodovsky, 1998, Wang et al., 2007), medical science (Yi and Beheshti, 2009, Tao et al., 2012) and ECG (electrocardiogram) analysis (Koski, 1996, Andreão et al., 2006).

Recent work has exploited the potential of the HMM to analyse the stock market and predict the financial market. Hassan and Nath (2005) apply HMMs to forecast airline stocks. They use the past datasets for the chosen airlines to train an HMM model and the trained HMM to search for the variable of interest behavioural data pattern in the past data. They forecast the airline stocks using the neighbouring values of these datasets. Gupta (2012) considers the fractional change in stock value and the intraday high and low values of the stock to train the continuous HMM and then uses this HMM to make a maximum a posteriori decision about all the possible stock values for the next day. Lee et al. (2014) use HMMs to learn the historical trend patterns of foreign exchange and to predict the next–day movement trends. Huang et al. (2015) extend selective HMMs to combine the financial index with the selected Twitter

mood to predict the next–day trends in the stock market. Fan et al. (2016) discuss the pricing of dynamic fund protection when the value process of the investment fund is governed by a geometric Brownian motion with parameters modulated by a continuous–time, finite state hidden Markov chain.

The other research direction of the HMM application is combining HMMs with other models. Hassan et al. (2007) propose a fusion model, combining an HMM with an ANN and a GA (genetic algorithms), to generate one–day–ahead forecasts for stock prices. In this model, the optimised HMM is used to identify similar data patterns from the historical data. Haeri et al. (2015) propose a hybrid approach using HMMs and classification and regression trees algorithms for forecasting the daily direction (the increase or decrease) of Euro–Yen exchange rates.

4.2.4 HMMs and three fundamental questions

An HMM can be described as $\lambda = \lambda(S, V, \pi, A, B)$ (Rabiner, 1989) , where:

1) $S = \{s_1, s_2 \ldots \ s_N\}$ is the set of states and N is the number of states in the model. Although the states are hidden, for many practical applications there is often some physical significance attached to the states or to sets of states in the model.
2) $V = \{v_1, v_2 \ \ldots \ v_M\}$ is the set of symbols. M is the number of observation symbols (which correspond to the physical output of the system being modelled).
3) $\pi = \{\pi_i\}$ for size N defines the initial probability distribution, $\pi_i = P(S(t = 0) = S_i)$, where π_i represents the probability of being in state i at the beginning of the experiment (i.e., at Time t = 0).

4) $A = \{a_{ij}\}$ for size N × N defines the transition matrix where $a_{ij} = P(s(t) = s_j | s(t-1) = s_i)$, the conditional probability from State i to State j.

5) $B = \{b_{ik}\}$ for size N × M defines the emission probability where $b_{ik} = P(O_t = v_k | s(t) = s_i)$, the probability of observing $O_t = v_k$ at State i.

To help understand the basic features of HMM, we present a simple example of the HMM applied to a hypothetical stock market in Figure 4.3. We assume that there are three states—good, bad and neutral—in the stock market. The arrows indicate allowed transitions. There are three types of observations: up, down and no change (NoC). In this example, $S = \{\text{Good}, \text{Neutral}, \text{Bad}\}$, $V = \{\text{Up}, \text{Down}, \text{No Change}\}$,

$$A = \begin{bmatrix} 0.4 & 0.4 & 0.2 \\ 0.3 & 0.4 & 0.3 \\ 0.2 & 0.4 & 0.4 \end{bmatrix}, \; B = \begin{bmatrix} 0.5 & 0.3 & 0.2 \\ 0.2 & 0.6 & 0.2 \\ 0.2 & 0.3 & 0.5 \end{bmatrix} \quad (4\text{–}8)$$

In this sample, we cannot observe the hidden states, that is we don't know the economic state is good, bad or neutral. Every day, the chance of the appearance of different hidden states follows the matrix A; the chance of the appearance of different observations follows the matrix B. We only can observe the different observations, that is, stock price increasing, decreasing or remaining stationary.

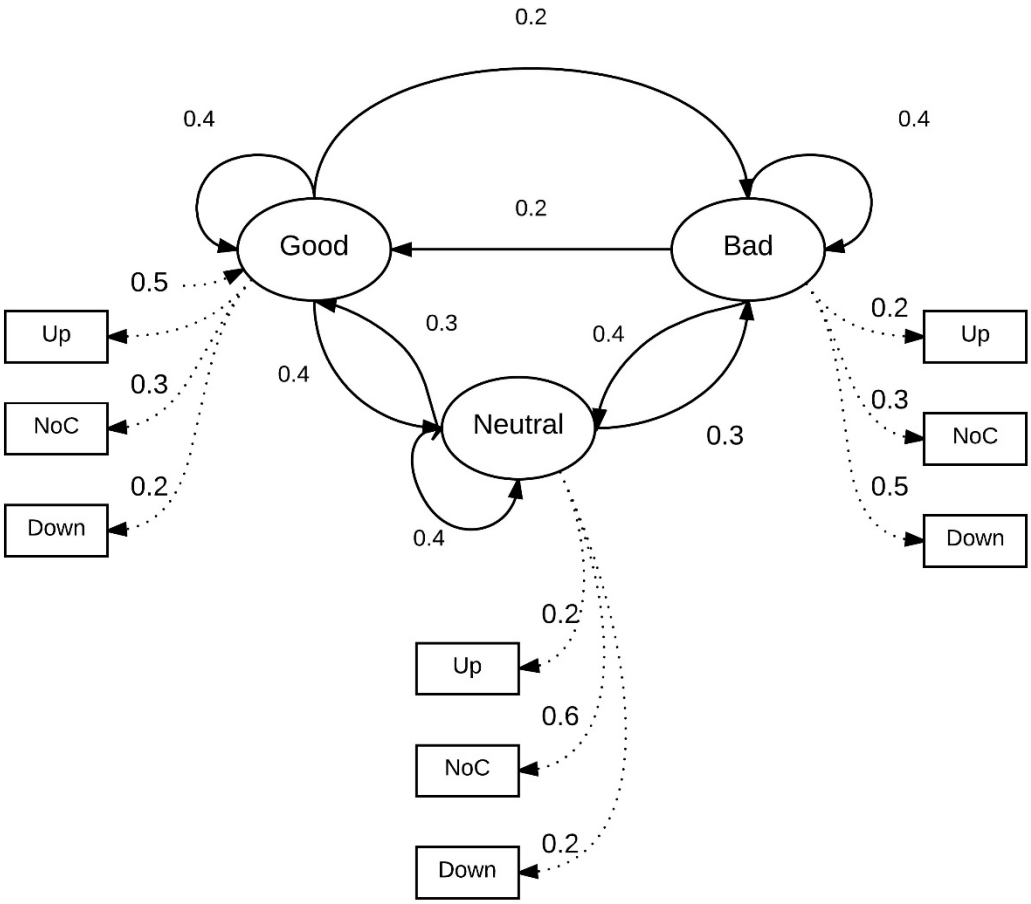

Figure 4.3 An example of the use of an HMM model to describe a hypothetical stock market

Each circle represents a hidden Markov state. There are three states—good, bad and neutral—in the hypothetical stock market. Arrows indicate allowed transitions. There are three types of observation: up, down and no change (NoC).

A hidden Markov model instantiates one assumption: the probability of an output observation O_1 depends only on the state that produced the observation q_i and not on any other states or any other observations, which can be called output independence.

$$P(o_i|q_1, q_2, \ldots, q_T, o_1, o_2, \ldots, o_T) = P(o_i|q_i) \quad (4\text{-}9)$$

To work with the HMM, the following three fundamental questions should be resolved:

1) Evaluation: given the model λ= (A, B, π), how do we compute P(O| λ), the probability of occurrence of the observation sequence $O = O_1, O_2 \ \ldots \ O_t$?

2) Decoding: given the observation sequence O and a model λ, how do we choose a state sequence $q_1,\ q_2\ \ldots\ q_t$. that best explains the observations?

3) Learning: given the observation sequence O and a space of models found by varying the model parameters A, B and π, how do we locate the model that best explains the observed data?

There are established algorithms to solve the above questions (Rabiner and Juang, 1986). The forward–backward algorithm to compute the P(O| λ) (Problem 1), the Viterbi algorithm to resolve Problem 2 and the Baum–Welch algorithm to train the HMM (Problem 3).

The forward–backward algorithm:

The forward–backward algorithm is based on the technique known as dynamic programming. Dynamic programming breaks a complex problem down into a collection of simpler sub–problems, solves each of these sub–problems only once, stores the solutions and uses them later, rather than recomputing them. The procedure of the forward–backward algorithm is shown below (Rabiner, 1989, Petrushin, 2000):

Let $\alpha_t(i) = P(O_{1,}O_2\ \ldots\ O_t, q_t = s_i|\lambda)$ be the probability of the partial observation sequence $O_{1,}O_2\ \ldots\ O_t$ to be produced by all possible state sequences that end at the state: $i-th$. The forward procedure is a recursive algorithm for calculating $\alpha_t(i)$ for the observation sequence of increasing length.

The Forward algorithm.

1. Initialisation: $\alpha_1(i) = \pi_i b_i(O_1) \qquad 1 \le i \le N$
2. Recursion: for $t = 2,3,\ldots,T$, and for $j = 1,2,\ldots\ ,N,$ compute

$$\alpha_t(j) = \left[\sum_{i\ =\ 1}^{N} \alpha_{t-1}(i)a_{ij}\right] b_j(O_t)$$

3. Termination: $P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$.

Similar to the forward algorithm, a symmetrical backward variable $\beta_t(i) = P(O_{t+1,}O_{t+2} \dots O_T|q_t = s_i, \lambda)$ is defined as the conditional probability of the partial observation sequence $(O_{t+1}, O_{t+2} \dots O_T)$ from $t+1$ to the end to be produced by all state sequences that start at the state: $i-th$. The backward procedure calculates recursively backward variables reversing along the observation sequence.

The Backward algorithm.

1. Initialisation: $\beta_{\mathrm{T}}(i) = 1 \qquad 1 \le i \le N$
2. Recursion: for $t = \mathrm{T}-1, \mathrm{T}-2, \dots, 1$, and for $i = 1,2, \dots, N$,

$$\beta_t(i) = \sum_{\mathrm{j}\ =\ 1}^{N} [a_{ij} b_j(O_{t+1})\beta_{t+1}(j)]$$

3. Termination: $P(O|\lambda) = \sum_{i=1}^{N} \pi_i \beta_1(i)$.

The Viterbi algorithm:

The Viterbi Algorithm, which could be interpreted as a dynamic programming algorithm, was first proposed by Andrew J. Viterbi in 1967 (Viterbi, 1967). The Viterbi algorithm chooses the state sequence that best maximises the likelihood of the state sequence for the given observation sequence. Let $\delta_t(i)$ be the maximal probability of state sequences of the length t that end in State i and produce the first t observations for the given model.

$$\delta_t(i) = \max_{q_1,q_2 \ldots q_{t-1}} P[q_1 q_2 \ldots q_{t-1}, q_t = i, O_1 O_2 \ldots O_t | \lambda] \quad (4\text{-}10)$$

The Viterbi algorithm.

1. Initialisation: $\delta_1(j) = \pi_j b_j(o_1)$**;** $\psi_1(j) = 0$

$$1 \le j \le N$$

2. Recursion: for $t = 2,3, \ldots, \mathrm{T}$, and for $j = 1,2, \ldots, N$,

$$\delta_t(j) = \max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}] b_j(O_t)$$

$$\psi_t(j) = \operatorname{argmax}_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}]$$

3. Termination: $P^* = \max_{1 \le i \le N} [\delta_T(i)]$

$$q_T^* = \operatorname{argmax}_{1 \le i \le N} [\delta_T(i)]$$

4. Backtracking: $q_t^* = \psi_{t+1}(q_{t+1}^*), \ t = T-1, T-2, \ldots, 1$

The Baum–Welch algorithm:

The Baum–Welch algorithm was proposed to estimate the parameters of the HMM model—that is, the initial probability distribution π, the transition probabilities A, and the emission functions B. The algorithm determines the locally optimal parameters by essentially using three equations: one for the initial probabilities π, one for the transition probabilities α_{ij} and one for the emission probabilities b_{ik}.

$$\pi_i = \frac{\mathrm{E}\ (\text{Number of times a sequence started with } S_i)}{\mathrm{E}\ (\text{Number of times a sequence started with any state})} \quad (4\text{–}11)$$

$$\alpha_{ij} = \frac{\mathrm{E}\ (\text{Number of times the state changed from } S_i \text{ to } S_j)}{\mathrm{E}\ (\text{Number of times the state changed from } S_i \text{ to any state})} \quad (4\text{–}12)$$

$$b_{ik} = \frac{\mathrm{E}(\text{Number of times the state was } S_i \text{ and the observation was } \mathrm{v}_k)}{\mathrm{E}\ (\text{Number of times the state was } S_i))} \quad (4\text{–}13)$$

These equations are used to recalculate the parameters of the model. The process continues until the stopping criterion is reached.

The Baum–Welch algorithm.

1. Initialisation: pre–set model parameters
2. Recursion: define $\xi_t(i,j) = P(q_t = i, q_{t+1} = j|O,\lambda)$ as the probability of moving from State i at t to State j at $t+1$; and define $\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j)$ as the probability of starting in State i at t. Based on the forward–backward algorithm,

$$\xi_t(i,j) = \frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}$$

   Then recalculate the model parameters π_i, α_{ij} and b_{ik}, for $1 \le i \le N$, $1 \le j \le N$,$1 \le k \le M$

$$\tilde{\pi}_i = \gamma_1(i)$$

$$\tilde{\alpha}_{ij} = \frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)}$$

$$\tilde{b}_{ik} = \frac{\sum_{t=1}^{T}(\gamma_t(i)\,|o_t = v_k)}{\sum_{t=1}^{T}\gamma_t(i)}$$

3. Termination: when the difference in the measure of the likelihood function between two consecutive iterations is less than the threshold or the maximal number of iterations is exceeded.

4.3 The proposed CMHMM model

4.3.1 The reason for proposing a new model

Compared with the successful applications of HMMs in engineering, applications of the regime–switching model and HMMs in finance are still being developed. Dacco and Satchell (1999) find that regime–switching models provide good in–sample performance, but that they are usually outperformed by random walks when used for forecasting (out–of–sample). They suggest

that the reason for this problem is that even if there is only a small misclassification, the forecasting of the regime–switching model will lose its advantage of knowing the correct model specification.

Conversely, for the application of HMMs to stock market prediction, most existing research (Hassan and Nath, 2005, Gupta, 2012, Huang et al., 2015) uses the stock price return or stock price for the observation series. However, in the HMM, the observation is identically and independently distributed, given the hidden state—a position that is obviously not true for stock returns. This is the disadvantage of using HMM to describe the stock market and the reason why we propose a new model.

4.3.2 The proposed CMHMM model

To forecast stock market indexes and stock prices, most existing research uses past price or past return, even where different models are used (Atsalakis and Valavanis, 2009). For instance, Göçken et al. (2016) build hybrid ANN models using technical indicators, such as the simple moving average of close price and the momentum of close price. Göçken et al. (2016) propose a forecasting model based on chaotic mapping, the firefly algorithm and support vector regression. To show the applicability of the proposed algorithm, they apply it to the daily closing stock prices of three NASDAQ firms. Ni et al. (2011) hybridise the fractal feature selection method and the support vector machine to predict the direction of the daily stock price index using past stock index prices.

In contrast, more recent research utilises news sentiment to predict stock price movement. For example, Li et al. (2014) propose a quantitative,

media–aware trading strategy to investigate the effect of media on stock markets. Liu et al. (2015) propose a model to both identify homogeneous stock groups and predict stock co–movement with firm–specific social media metrics. Ho and Wang (2016) propose neural network models using news sentiment to predict the stock price movement of GOOG.

The unique characteristic of HMMs is that the underlying system state is not directly observable and can only be estimated using related observable parameters. In this research, we can consider the stock price or return and the news sentiment as the observation series. Conversely, Wang and Wang (2017) build a game theoretical model to examine how the information advantage of insiders affects stock price movements. In their model, they define a variable called 'economic state' and assume: (1) the economic state can only be changed by the occurrence of a news event; (2) each role in the stock market cannot precisely know the previous economic state, the current economic state or the future economic state; and (3) the price of the stock is affected by the current economic state.

Inspired by this model, we consider the different levels of 'economic state' as hidden states in the HMM. Each 'economic state' has a significant chance to generate different levels news sentiment and different levels stock return. We can observe stock return and news sentiment to estimate the hidden state, as the stock returns are not independent in the days following news releases. We therefore consider the return to be affected by the past stock return (a Markov model) and an HMM model.

Figure 4.3 shows the general structure of a CMHMM model, which includes a discrete HMM and a Markov model. $\mathrm{Here}, s_1, s_2 \ldots\ s_t$ is the hidden

states series and $O_{11}, O_{12} \ldots O_{1t}$ is the Observation Series 1. This observation series is only affected by the hidden state s_i. $O_{21}, O_{22} \ldots O_{2t}$ is the Observation Series 2, while $O_{H1}, O_{H2} \ldots O_{Ht}$ is the effect of the HMM model and $O_{M1}, O_{M2} \ldots O_{Mt}$ is the effect of the Markov model.

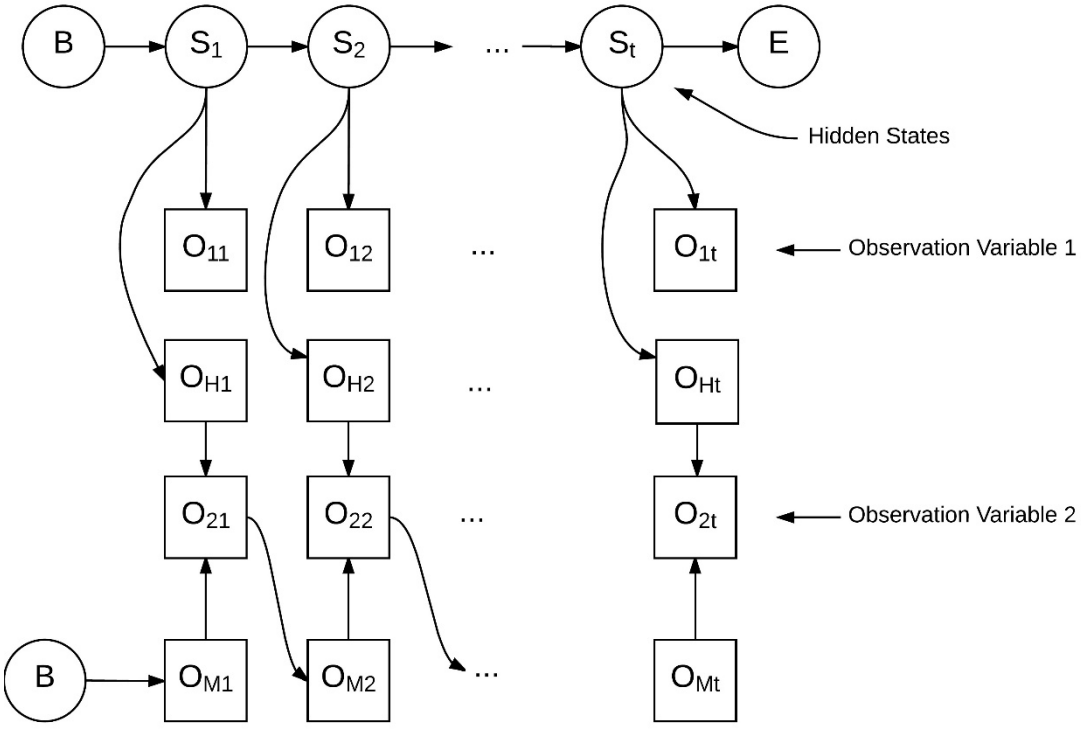

Figure 4.4 General Structure of a CMHMM model

For the purpose of illustration, we consider a financial market in which there exist different levels of economic states (for instance, three levels of economic states: good, bad and neutral) that are hidden and cannot be observed. However, news events and stock price changes can be observed. A favourable economic state means that the economic environment has a positive effect on the news sentiment and the stock price during the studied investment period, while an unfavourable economic factor means a negative

effect. The stock price change is not only affected by the hidden economic state; it is also affected by past fluctuations in stock price.

A CMHMM model can be described as $\lambda = \lambda(S_H, V1, V2, \pi1, A1, B1, B2, \pi2, A2)$, where:

1) $S_H = \{s_{H1}, s_{H2} \ldots s_{HN}\}$ is the set of hidden states, while N is the number of states in the model.

2) $V1 = \{v1_1, v1_2 \ldots v_{1M}\}$ is the set of symbols. M is the number of observation symbols that corresponds to the physical output observation series O_{1t} of the HMM.

3) $V2 = \{v2_1, v2_2 \ldots v2_Q\}$ is the set of symbols. Q is the number of observation symbols that corresponds to the physical output observation series O_{2t} of CMHMM, the output of series O_{Ht} of the HMM and the output of series O_{Mt} of the Markov model. For ease of understanding, we use $S_M = \{s_{M1}, s_{M2} \ldots s_{MQ}\}$ to describe states in the Markov model. $S_M = V2$.

4) $\pi1 = \{\pi1_i\}$ for size N defines the initial probability distribution of HMM, $\pi1_i = P(S_H(t = 0) = S_{Hi})$, where $\pi1_i$ represent the probability of being in State i at the beginning of the experiment (i.e., at Time t = 0).

5) $A1 = \{a1_{ij}\}$ for size N × N defines the transition matrix of HMM with $a1_{ij} = P(s_H(t) = s_{Hj} | s_H(t-1) = s_{Hi})$ the conditional probability from HMM hidden State i to hidden State j.

6) $B1 = \{b1_{ik}\}$ of size N × M defines the emission probability with $b1_{ik} = P(O_1(t) = v1_k | s_H(t) = s_{Hi})$, the probability of observing O_{1t} at State i.

7) $B2 = \{b2_{il}\}$ of size N × Q defines the emission probability with $b2_{il} = P(O_H(t) = v2_l | s_H(t) = s_{Hi})$, the probability of observing O_{Ht} at State i.

8) $\pi 2 = \{\pi 2_i\}$ for size Q defines the initial probability distribution of Markov model, $\pi 2_i = P(S_M(t = 0) = S_{Mi})$, where $\pi 2_i$ represents the probability of being in State i at the beginning of the experiment (i.e., at Time t = 0).

9) $A2 = \{a2_{ij}\}$ for size Q × Q defines the transition matrix of the Markov model with $a2_{ij} = P(s_M(t) = s_{Mj} | s_M(t-1) = s_{Mi})$, the conditional probability from Markov model State i to State j.

Additionally, there are two observation series, O_{1t} and O_{2t}; O_{1t} is only affected by the HMM part, whereas O_{2t} is affected by the HMM part and the Markov model part. Here, O_{Ht} describes the effect of the HMM part, while O_{Mt} denotes the effect of the Markov model part.

4.3.3 Three fundamental questions for CMHMM

Just as for the HMM model, to work with CMHMM, the following three fundamental questions should be resolved:

1. Evaluation: given the model λ= (A1, A2, B1, B2, π1, π2) how do we compute P(O|λ), the probability of occurrence of the observation sequence $O = \begin{bmatrix} O_{11,} O_{12} \dots O_{1T} \\ O_{21}, O_{22} \dots O_{2T} \end{bmatrix}$.
2. Decoding: given the observation sequence O and a model λ, how do we choose a state sequence $\begin{bmatrix} q_{11,} q_{12} \dots q_{1T} \\ q_{21}, q_{22} \dots q_{2T} \end{bmatrix}$ that best explains the

observations. Here, $[q_{11,}q_{12,}\dots q_{1T}]$ is the state sequence of the HMM and $[q_{21}, q_{22}, \dots q_{2T}]$ is the state sequence of the Markov model.

3. Learning: given a model and the observation sequence $O = \begin{bmatrix} O_{11,}O_{12} \dots O_{1T} \\ O_{21}, O_{22} \dots O_{2T} \end{bmatrix}$, how do we adjust the model parameters (A1, A2, B1, B2, π1, π2) to best explain the observed data.

In the CMHMM mode, the O_{2t} is affected by the HMM part and the Markov model part, where O_{Ht} describes the effect of the HMM model part and O_{Mt} denotes the effect of the Markov model part. We will answer the above questions using the revised forward–backward algorithm to compute the P(O|λ) (evaluation problem), the revised Viterbi algorithm to resolve the decoding problem and the revised Baum–Welch algorithm to address the learning problem.

The revised forward–backward algorithm for CMHMM:

We follow the idea of breaking a complex problem down into a collection of simpler sub–problems, to solve each of these sub–problems only once and to store the solutions for later use, rather than recomputing them. For our CMHMM model, the revised forward–backward algorithm is shown below:

Let $\alpha_t(i) = P(\begin{bmatrix} O_{11,}O_{12} \dots O_{1t} \\ O_{21}, O_{22} \dots O_{2t} \end{bmatrix}, q_{Ht} = s_{Hi}|\lambda)$ be the probability of the partial observation sequence $\begin{bmatrix} O_{11,}O_{12} \dots O_{1t} \\ O_{21}, O_{22} \dots O_{2t} \end{bmatrix}$ to be produced by all possible state sequences that end at State $i-th$ of HMM at Time t. In the CMHMM model, O_{2t} is the observation and the state of the Markov model at Time t. The forward procedure is a recursive algorithm for calculating $\alpha_t(i)$ for the observation sequence of increasing length. The first step is Initialisation; we

calculate the $\alpha_t(i)$ when t = 1. The $\alpha_1(i)$ is the joint probability of the event; the Observation 1 is O_{11} and the Observation 2 is O_{21}. These two events are independent. The probability of the first event is $\pi 1_i b1_i(O_{11})$. The probability of the second event, given that the hidden state is at s_i, is the sum probability of $[b2_i(O_{\mathrm{H1}})] * \pi 2_{O_{M1}}$ when $O_{H1} + O_{M1} = O_{21}$. Similarly, we can calculate the $\alpha_t(i)$ when $t = 2,3 \ \ldots \ T$.

The revised forward algorithm for CMHMM.

1. Initialization:

$$\alpha_1(i) = \pi 1_i b1_i(O_{11}) * \sum_{when\ O_{H1}+O_{M1}=O_{21}} \{[b2_i(O_{\mathrm{H1}})] * \pi 2_{O_{M1}}\}$$

$$1 \le i \le N$$

2. Recursion: for $t = 2,3, \ldots, T$, and for $i = 1,2, \ldots, N$,

$$\alpha_t(i) = \{[\sum_{k=1}^{N} \alpha_{t-1}(k)\alpha 1_{ki}]b1_i(O_{1t})\} * \sum_{when\ O_{Ht}+O_{Mt}=O_{2t}} \{b2_i(O_{Ht}) * a2(O_{2,(t-1)}, O_{Mt}) \ \}.$$

3. Termination: $P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$.

Similarly, a symmetrical backward variable $\beta_t(i) = P(\begin{bmatrix} O_{1,t\ +\ 1,} O_{1,t\ +\ 2} \ldots O_{1,T} \\ O_{2,t\ +\ 1}, O_{2\ ,t\ +\ 2} \ldots O_{2,T} \end{bmatrix} | q_{Ht} = i, \lambda)$ is defined as the conditional probability of the partial observation sequence $\begin{bmatrix} O_{1,t\ +\ 1,} O_{1,t\ +\ 2} \ldots O_{1,T} \\ O_{2,t\ +\ 1}, O_{2,t\ +\ 2} \ldots O_{2,T} \end{bmatrix}$ from $t + 1$ to the end, to be produced by all state sequences that start at State $i - th$. The backward procedure calculates recursively backward variables reversing throughout the observation sequence.

The revised backward algorithm for CMHMM.

1. Initialization: $\beta_{\mathrm{T}}(i) = 1 \qquad 1 \le i \le N$
2. Recursion: for $t = \mathrm{T}-1, \mathrm{T}-2, \ldots, 1$, and for $i = 1,2, \ldots, N$,

   $\beta_t(i) = \sum_{\mathrm{k}=1}^{N}\{\alpha 1_{ik} b1_k\left(O_{1,(t+1)}\right)\beta_{t+1}(k) *$

   $\sum_{when\ O_{H(t+1)}+O_{M(t+1)}=O_{2(t+1)}}[b2_k\left(O_{2,H(t+1)}\right) * a2\left(O_{2t}, O_{M(t+1)}\right)\}$.
3. Termination:

   $P(O|\lambda) = \sum_{i=1}^{N}\{\pi 1_i \beta_1(i) * \sum_{when\ O_{H1}+O_{M1}=O_{21}}\{[b2_i(O_{\mathrm{H1}})] * \pi 2_{O_{M1}}\}\}.$

The revised Viterbi algorithm for CMHMM:

The revised Viterbi algorithm chooses the state sequence that best maximises the likelihood of the state sequence for the given observation sequence. In the CMHMM, we have two state sequences: the state sequence of the HMM part $q_{11}, q_{12} \ldots q_{1T}$ and the state sequence of the Markov model part $q_{M1}, q_{M2}, \ldots, q_{MT}$ Let $\delta_t(i,j)$ be the maximal probability of state sequences for the length t that end the HMM part in State i and the HMM part in State j and produce the first t observations for the given model:

$$\delta_t(i,j) \qquad (4\text{-}14)$$

$$= \max_{\begin{bmatrix} q_{11}, \ldots q_{1(t-1)} \\ q_{M1}, \ldots q_{M(t-1)} \end{bmatrix}}\{P\left[\begin{bmatrix} q_{11}, \ldots q_{1(t-1)} \\ q_{M1}, \ldots q_{M(t-1)} \end{bmatrix}, q_{1t} = i, q_{Mt} = j, \begin{bmatrix} O_{11,} \ldots O_{1t} \\ O_{21}, \ldots O_{2t} \end{bmatrix} \middle| \lambda\right]\}$$

The revised Viterbi algorithm for CMHMM.

1. Initialization:

   $\delta_1(i,j) = \pi 1_i b1_i(o_{11})^* \sum_{when\ O_{H1}+S_{2j}=O_{21}}\{b2_i(O_{\mathrm{H1}}) * \pi 2_j\};$
   $\psi_1(i,j) = 0$

$1 \le i \le N,\ 1 \le j \le Q$

2. Recursion: for $t = 2,3, \ldots, \mathrm{T}$, and for $i = 1,2, \ldots, N;\ j = 1,2, \ldots, \mathrm{Q}$

$$\delta_t(i,j) = max_{k,l}\{[\delta_{t-1}(k,l)a1_{ki}]b1_i(o_{1t}) * \sum_{when\ O_{Ht}+S_{2j}=O_{2t}} \left\{b2_i(O_{Ht}) * a2\left(O_{2,(t-1)}, S_{2j}\right)\right\}\}$$

$$\psi_t(i,j) = \mathrm{argmax}_{k,l}(\delta_t(i,j))$$

3. Termination: $P^* = \max_{1 \le i \le N, 1 \le j \le Q}\{\delta_{\mathrm{T}}(i,j)\}$

$$q_T^* = \mathrm{argmax}_{1 \le i \le N}[\delta_T(i,j)]$$

4. Backtracking: $q_t^* = \psi_{t+1}(q_{t+1}^*)\ \ t = T-1, T-2, \ldots, 1$

The revised Baum–Welch algorithm for CMHMM:

In the CMHMM model, we have the HMM part and the Markov model parts. We revise the Baum–Welch algorithm to estimate the parameters of the HMM model part and the parameters of the Markov model part. For the HMM model part, the revised Baum–Welch algorithm determines the locally optimal parameters by essentially using four equations: one for the initial probabilities $\pi 1$, one for the transition probabilities $\alpha 1_{ij}$, one for the emission probabilities $b1_{ik}$ and one for the emission probabilities $b2_{il}$.

$$\pi 1_i = \frac{\mathrm{E}\ (\text{Number of times a sequence started with } \mathrm{s}_{Hi})}{\mathrm{E}\ (\text{Number of times a sequence started with any state})} \quad (4\text{–}14)$$

$$\alpha 1_{ij} = \frac{\mathrm{E}\ (\text{Number of times the state changed from } \mathrm{s}_{Hi} \text{ to } \mathrm{s}_{Hj})}{\mathrm{E}\ (\text{Number of times the state changed from } \mathrm{s}_{Hi} \text{ to any state})} \quad (4\text{–}15)$$

$$b1_{ik} = \frac{\mathrm{E}(\text{Number of times the state was } \mathrm{s}_{Hi} \text{ and the observation was }}{\mathrm{E}\ (\text{Number of times the state was } \mathrm{s}_{Hi}))} \quad (4\text{–}16)$$

$$b2_{il} = \frac{\text{E(Number of times the state was } s_{Hi} \text{ and the observation was}}{\text{E (Number of times the state was } s_{Hi}))} \quad (4\text{–}17)$$

For the Markov model part, the locally optimal parameters are determined by essentially using two equations: one for the initial probabilities $\pi 2$ and one for the transition probabilities $\alpha 2_{ij}$.

$$\pi 2_i = \frac{\text{E (Number of times a sequence started with } s_{Mi})}{\text{E (Number of times a sequence started with any state)}} \quad (4\text{–}18)$$

$$\alpha 2_{ij} = \frac{\text{E (Number of times the state changed from } s_{Mi} \text{ to } s_{Mj})}{\text{E (Number of times the state changed from } s_{Mi} \text{ to any state)}} \quad (4\text{–}19)$$

These six equations are used to recalculate the parameters of the model. The process continues until the stopping criterion has been reached.

The revised Baum–Welch Algorithm for CMHMM.

1. Initialization: pre–set model parameters
2. Recursion: for HMM part, we define $\xi_t(i,j) = P(q_{Ht} = S_{Hi}, q_{H(t+1)} = S_{Hj}|O,\lambda)$ as the probability of moving from state S_{Hi} at t to S_{Hj} at $t+1$; $\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j)$ as the probability of starting in S_{Hi} at t. We note $P2 = \sum_{when\ O_{H(t+1)}+O_{M(t+1)}=O_{2(t+1)}} [b2_i(O_{2,H(t+1)}) * a2(O_{2t}, O_{M(t+1)})]$. Based on the forward–backward algorithm,

$$\xi_t(i,j) = \frac{\alpha_t(i)[a1_{ij} b1_j(O_{1(t+1)}) * P2]\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_t(i) a1_{ij} b1_j(O_{1(t+1)} * P2)\beta_{t+1}(j)}$$

Then we recalculate the model parameters $\pi 1_i$, $\alpha 1_{ij}$, $b1_{jk}$ and $b2_{jk}$.

$$\widetilde{\pi 1_\iota} = \gamma_1(i)$$

$$\widetilde{\alpha 1}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$$

$$\tilde{b}1_{jk} = \frac{\sum_{t=1}^{T} (\gamma_t(j) \,|O1_t = v1_k)}{\sum_{t=1}^{T} \gamma_t(j)}$$

$$\tilde{b}2_{jk} = \frac{\sum_{t=1}^{T} (\gamma_t(j) \,|O_{Ht} = v2_k)}{\sum_{t=1}^{T} \gamma_t(j)}$$

For the HMM part, we define $\delta(i,j) = P(q_{Mt} = S_{Mi}, q_{M(t+1)} = S_{Mj}|O,\lambda)$ as the probability of moving from state S_{Mi} at t to S_{Mj} at $t+1$; $\delta_t(i) = \sum_{j=1}^{N} \delta_t(i,j)$ as the probability of starting in S_{Mi} at t. Then recalculate the model parameters $\pi 2_i$, $\alpha 2_{ij}$

$$\widetilde{\pi 2}_\iota = \delta_1(i)$$

$$\widetilde{\alpha 2}_{ij} = \frac{\sum_{t=1}^{T-1} \delta_t(i,j)}{\sum_{t=1}^{T-1} \delta_t(i)}$$

3. Termination: when the difference in the measure of the likelihood function between two consecutive iterations is less than a threshold or the maximal number of iterations is exceeded.

The Baum–Welch Algorithm and the revised Baum–Welch Algorithm for CMHMM are special cases of the Expectation Maximization (EM) algorithm (Bilmes, 1998). The EM algorithm (Dempster et al., 1977) is an iterative method of finding the maximum–likelihood estimate of parameters in statistical models from a given data set when there are unobserved latent variables in the model.

The EM iteration alternates between performing an expectation (E) step and a maximization (M) step. The E step finds the expectation of the log–likelihood evaluated with respect to the unknown data given the observed data and the current parameter estimates. The M step finds parameters that maximise the expected log–likelihood. Each iteration is

guaranteed to increase the log–likelihood and the algorithm is guaranteed to converge to a local maximum of the likelihood function.

4.4 Applications of the proposed CMHMM model in stock price prediction

4.4.1 Data

For this research, we consider only one and a half years (Jan 2015–Jun 2016) of stock price returns and news items sentiment. For the stock price, we examine the DJIA index. Figure 4.4 shows the daily closing prices of the DJIA index during the period from January 2015 to June 2016.

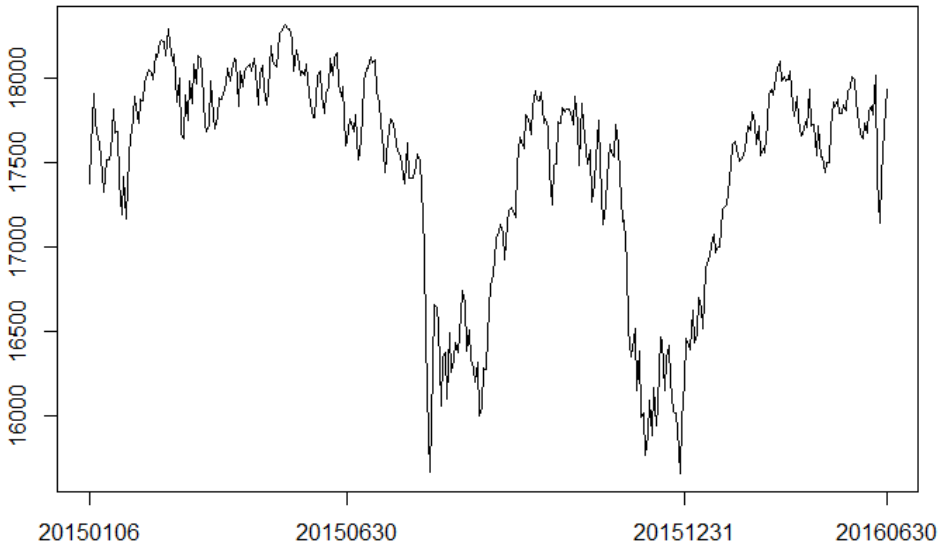

Figure 4.5 Daily closing prices of the DJIA index (Jan 2015–Jun 2016)

Our raw news items data were obtained from the RPNA database (see Appendix for further details). The database contains a unique observation for

every article and includes the date and time at which each news article was released, a unique firm identifier and several variables that quantify the content and form of each article.

There are 39 fields that are used to describe each news item. For this research, we only consider only some of these fields, such as the time stamp, the company name, the relevance of the news, the event sentiment and the novelty of the news. The ‘relevance’ score ranging from 0 to 100 indicates how strongly a news story is related to the entity under examination and a score of 100 suggests the article is highly relevant. For a news story with a relevance score of 100, the ENS represents its novelty value; the first story reporting a categorised event receives a novelty score of 100. The ESS represents the news sentiment for a given entity, ranging from 0 to 100, where 0 indicates extremely negative news, 50, neutral news and 100, extremely positive news.

We have ascertained the relevance of the news to the 29 stocks based on the field ‘company name’ and summarised the statistics of the daily news volume and news sentiment. To analyse the effect of the news on the stock market, we describe the news that happens before the stock market’s opening on Day i as the news of Day $i-1$. To build the daily news volume and news sentiment series, we need to consider the market hours of the NASDAQ stock market and the NYSE, which run from 9:30a.m. to 4:00p.m. Further, we need to consider summer daylight–savings time to pre–process our data set, as 2:00am on 9 March 2014, 8 March 2015 and 13 March 2016 will become 3:00am and 2:00am on 2 November 2014 and 1 November 2015 will become 1:00am.

For each news item, the ESS represents the news sentiment for a given entity, ranging from 0 to 100, where 0 indicates extremely negative news, 50 indicates neutral news and 100 indicates extremely positive news. To easily understand the effect of the news, we use –50 to indicate extremely negative news, 0 to designate neutral news and +50 to connote extremely positive news for each item. We only consider all news sentiment of the DJIA components. Table 4.1 shows the basic statistics for the daily total news sentiment of DJIA and DJIA components and Figure 4.5 describe daily total sentiment of the DJIA components during Jan 2015 to Jun 2016.

Table 4. 1 The basic statistics for the daily total news sentiment (Jan 2015–Jun 2016)

	Symbol	**Mean**	**Median**	**Max**	**Min**	**Std. dev.**
	DJIA	95.81	84	509	–107	90.51
1	MMM	1.90	0	119	–39	12.54
2	DD	1.70	0	146	–41	12.40
3	MCD	1.56	0	148	–44	15.45
4	XOM	1.38	0	54	–57	11.55
5	MRK	3.74	0	108	–31	12.24
6	AXP	1.82	0	120	–62	13.45
7	GE	15.27	0	166	–32	27.37
8	MSFT	4.85	0	87	–32	14.48
9	PFE	5.31	0	180	–56	19.50
10	HD	3.14	0	194	–39	20.45
11	PG	1.03	0	158	–45	15.06
12	BA	7.31	0	228	–56	23.56
13	INTC	2.93	0	87	–81	14.67
14	TRV	–1.24	0	13	–43	4.44
15	CAT	–0.57	0	130	–80	12.95
16	IBM	7.78	0	117	–65	15.94
17	UTX	4.38	0	123	–62	17.66
18	CVX	1.22	0	66	–134	12.82
19	JNJ	4.03	0	135	–87	18.05

20	VZ	3.93	0	196	–52	18.46
21	CSCO	3.79	0	173	–47	19.51
22	JPM	4.47	0	142	–59	16.38
23	WMT	1.07	0	113	–51	14.82
24	KO	3.92	0	179	–51	19.09
25	DIS	2.42	0	109	–65	14.02
26	UNH	2.16	0	149	–43	15.37
27	GS	2.72	0	236	–121	19.99
28	NKE	1.01	0	145	–196	20.19
29	V	2.76	0	182	–57	18.02

This table presents the summary descriptive statistics for the daily total news sentiment (from Jan 2015 to Jun 2016) for the 29 stocks used in this study. The summary statistics include mean value (Mean), median value (Median), maximum (Max), minimum (Min) and standard deviation (Std. dev.).

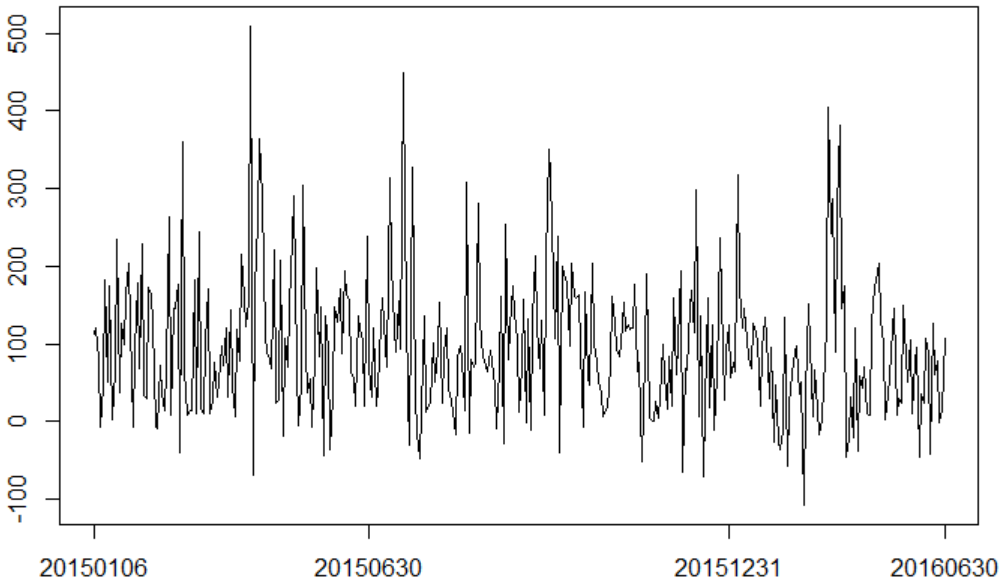

Figure 4.6 Daily total sentiment of the DJIA components (Jan 2015–Jun 2016)

4.4.2 The prediction method based on CMHMM

In our CMHMM model, we have two observation series; one is O_{1t}, which corresponds to the set of symbols $V1 = \{v1_1,\ v1_2\ \ldots v_{1M}\}$; the other is O_{2t}, which corresponds to the set of symbols $V2 = \{v2_1, v2_2\ \ldots v2_Q\}$. Using the prediction method based on CMHMM, we consider the news sentiment for the Observation Series O_{1t} and the log stock returns for the Observation Series O_{1t}. The first step using the CMHMM is to discretise the news sentiment series and the stock returns series. After this, we can use the data to produce a trained CMHMM model or to discern the probability of occurrence for the observation sequence.

During the training procedure for the CMHMM model, the parameters of the model $\lambda(S, V1, V2, \pi1, A1, B1, B2, \pi2, A2)$ are adjusted to maximise the probability that the given observation sequence O will be generated by the model. This is called the learning problem and was solved by the revised Baum–Welch algorithm. During the test period, the revised forward–backward algorithm is used to calculate probability.

Figure 4.6 shows the training period wherein CMHMM is used for prediction. We form the initial parameters of the CMHMM model and then update these parameters using the revised Baum–Welch algorithm based on the discretised news sentiment and stock returns. When the algorithm converges, a trained CMHMM model is produced. Using different test data, we can obtain a number of trained CMHMM models.

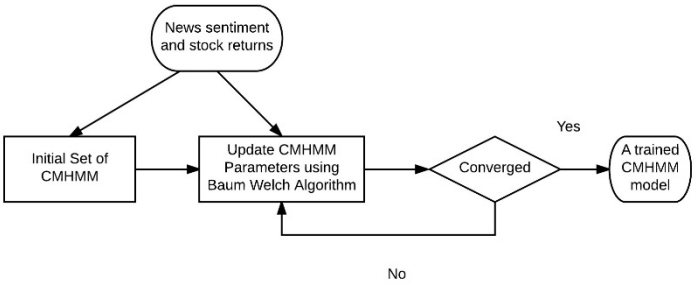

Figure 4.6 The training period using CMHMM for prediction

Figure 4.7 shows the test period wherein CMHMM is used for prediction. For a given, discretised news sentiment and stock returns series, the probability of the occurrence of these observation sequences is calculated by the revised forward–backward algorithm for all trained CMHMM models. The model with the largest $P(O|\lambda)$ is selected. The raw data employed to train this CMHMM model are used to predict the test data. That is, we consider the test data with the same future movement as the raw data which are used to obtain the largest $P(O|\lambda)$.

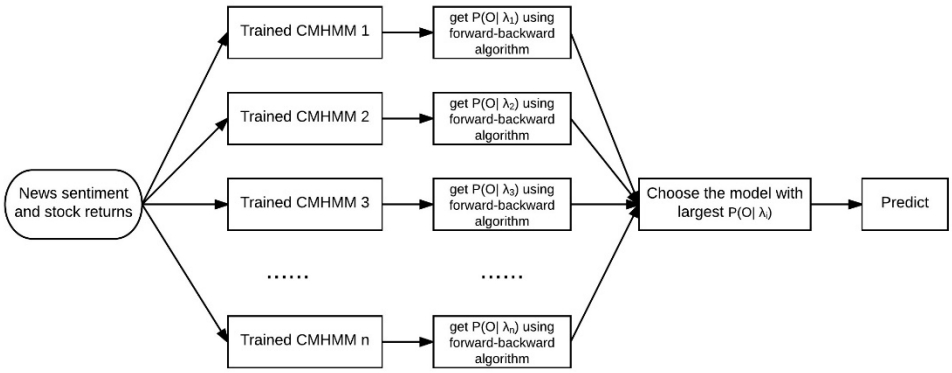

Figure 4.7 The test period using CMHMM for prediction

4.4.3 Empirical results

The data set has been divided into a training set (spanning one year from 2 January 2015 to 31 December 2015) and a test set (spanning six months from 2 January 2016 to 30 June 2016). We use the training set to build the model and the test set to evaluate the trading strategy.

For the HMM algorithms, there are several existing general–purpose software programmes, such as MATLAB and R, which are implementations of the various HMM algorithms. However, for our CMHMM model, it is necessary to code functions to implement the algorithms. In our simulations, we utilise MATLAB for coding.

The first step of our simulation is discretising the news sentiment series and the stock returns series. For the news sentiment, we calculate the daily $\sum(\mathrm{ESS} - 50)$ for the 29 DJIA index components and produce the daily news sentiment series. (We consider the news that reported public holidays as having occurred the previous weekday.) We use 20 as the cut-off for discretising this news sentiment series—that is, we classify all daily total news sentiment with a value of less than –20 as Group 1, daily total news sentiment between –20 and 20 ($-20 \leq$ daily total news sentiment ≤ 20) as Group 2 and daily total news sentiment larger than 20 as Group 3. We then obtain a discretised daily news sentiment series; for example, {2 1 3 3 2 1 3}.

For the stock returns, we use the DJIA opening prices to produce the daily log returns series $\mathrm{r_i}$, defined as $\mathrm{r_i} = \ln(\frac{p_{i\,+\,1}}{p_i})$. We use 0.01 as the cut-off for discretising this return series—that is, we classify daily log returns with values of less than –0.01 as Group 1, daily log returns between –0.01 and 0.01 (–0.01 $\leq$ $d$aily log return ≤ 0.01) as Group 2 and daily log returns larger

than 0.01 as Group 3. This observation series O_2 is the sum effect of O_{Hi} and O_{Mi}; $O_{H1}, O_{H2} \ldots O_{Ht}$ is the effect of the HMM model and $O_{M1}, O_{M2} \ldots O_{Mt}$ is the effect of the Markov model. We define the sum operations between O_H and O_M as shown in Table 4.2.

Table 4. 2 The basic operations between O_H and O_M

O_H	O_M	O_2
Group 1	Group 1	Group 1
Group 1	Group 2	Group 1
Group 1	Group 3	Group 2
Group 2	Group 1	Group 1
Group 2	Group 2	Group 2
Group 2	Group 3	Group 3
Group 3	Group 1	Group 2
Group 3	Group 2	Group 3
Group 3	Group 3	Group 3

This table presents the summary operations between O_H and O_M. O_H, O_M are input; O_2 is the output.

Figure 4.8 shows the trading strategy based on CMHMM. We assume we have fixed money to buy stock (That is the DJIA index). We only can buy stock or sell stock; we cannot borrow money or short stock. Before the opening of the market, we use CMHMM to predict the direction of the DJIA index movement. If the movement is uptrend and the asset status is 0 (hold money), we buy the stock. If it is downtrend and the asset status is 1 (hold stock), we sell the stock. We also update the asset status. If we hold stock, the asset status is 1; if we hold money, the asset status is 0.

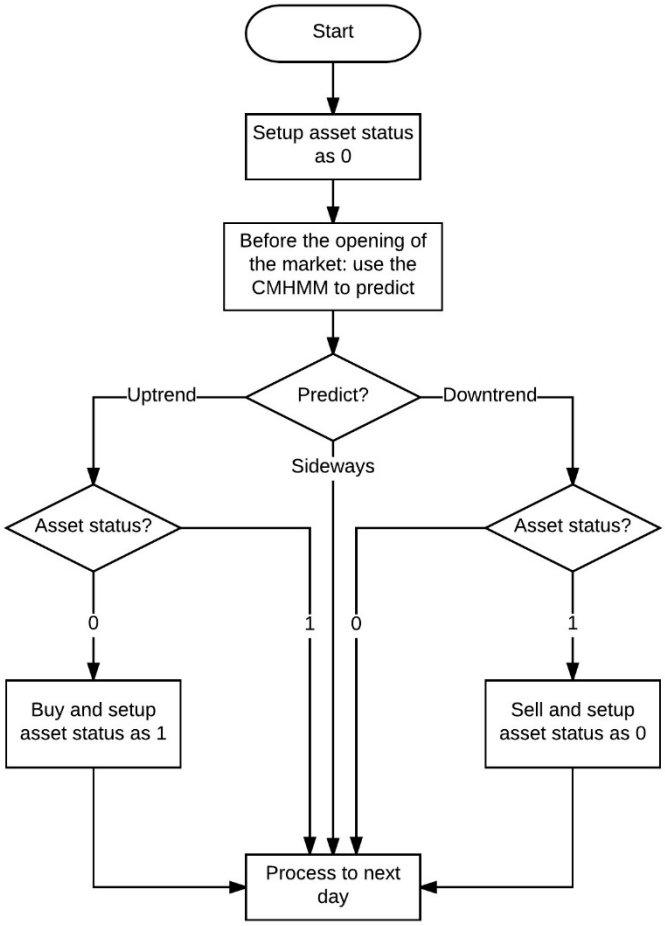

Figure 4.18 The trading strategy using CMHMM

Table 4.3 shows the performance of the proposed trading strategies. We use different observation lengths (from 16 to 20) to train the CMHMM model and assess the proposed trading strategies. The best performance is when the observation length is 18 and the return is 10.23%. The worst performance is when the observation length is 16 and the return is 5.05%. However, all of these performances are better than those of the DJIA index, which returns 2.98% during the test period.

Table 4.3 The performance characteristics of the CMHMM prediction model

Observation length	**Return (%)**
16	5.05
17	3.40
18	10.23
19	6.30
20	7.10
DJIA index	2.98

4.5 Conclusion

The goal of this research is to develop methods for modelling and forecasting stock market data. One of the main reasons for this is that most existing applications of HMMs in finance use stock returns or stock prices for the observations, assuming that they are independent in accordance with the requirements of the HMM model. In this chapter, we have proposed a new model (CMHMM), in which the observation is conducted using a Markov model and an HMM model. The new model provides a flexible, general–purpose approach for modelling various dynamic systems that can be observed through univariate or multivariate time series. We have discussed the evaluation, decoding and learning problems associated with the CMHMM as well as the application of the CMHMM, whereby news sentiment functions as one observation and the stock return as the other. The proposed model adheres to the nature of the stock market. The empirical results of the trading strategy provided by the CMHMM show the potential applications of the proposed model.

For this research, we have used only news sentiment and stock prices for the observation series; however, other time series, such as those produced by social media, may be considered. In future research, we will consider such data to improve our model.

Chapter 5:
Conclusions and future works

5.1 Conclusions

Our implicit goal in this thesis is to bridge the gap between academic and practical approaches by proposing methods and procedures that are theoretically sound and, at the same time, easily accessible for stock market prediction. In addition to exploring nonlinear model applications in the stock market, we have sought to investigate the relationship between news and the stock market.

In this thesis, we discussed the stock market prediction abilities of the ANN model, the kNN algorithm and our proposed CMHMM model. The empirical results, using past stock prices and news sentiment, show the potential trading methods based on each of these models.

Further, the empirical results from this dissertation contribute to an understanding of the relationship between news and the stock market. For instance, we find that news volume is not the Granger cause of stock price change, but that it contributes to portfolio variance both in– and out–of–sample; conversely, news sentiment is the Granger cause of stock price change and positive news sentiment contributes to the portfolio return in–sample, while negative news sentiment contributes to the portfolio return out–of–sample—as a consequence of investor overreaction to it.

5.2 Future works

In this dissertation, we have discussed the relationship between news and the stock market and explored methods for forecasting stock price movement and portfolio optimisation using nonlinear models based on past stock prices and news items. However, there is still significant work that can be done to accurately predict stock price movement.

5.2.1 Forecasting financial market movement with state–space models

State–space models were first used in the control theory for the modelling of continuously changing, unobserved state variables, which may be estimated by the Kalman filter. The Kalman filter algorithm plays a central role in the modelling and can also be used to estimate and further predict the states of state–space models.

Recently, scholars have extended the state–space model in the domain of economics and finance. For instance, Balke and Wohar (2002) employ a state–space model to explore the dynamics of the log price–dividend ratio alongside long– and short–term interest rates, real dividend growth and inflation. They find that the advantage of the state–space approaches is that they can parsimoniously model the low–frequency movements present in the data. Al-Anaswah and Wilfling (2011) use a state–space model with Markov switching to detect speculative bubbles in the financial market. They estimate a two–regime Markov–switching specification for the unobservable bubble process, which includes a scenario in which the bubble survives and one in

which it collapses. Škovránek et al. (2012) present a macroeconomic model to investigate the behaviour of the national economies of the three Commonwealth countries. The model, based on state–space modelling, uses state variables to describe the behaviour of a system, the gross domestic product, inflation and the unemployment rate.

In future studies, we will construct a model of news sentiment and stock prices in state–space form, which we will estimate using the Kalman filter. In the empirical analysis, we will apply our methodology to real–world datasets and predict stock price movement.

5.2.2 Analysis of high–frequency financial data using non–linear models

In Essays 1 and 3, we considered daily news information and stock returns and, in Essay 2, we utilised weekly data to build a trading strategy or portfolio. The main reason that we used daily/weekly data is that we needed to decrease the trade number to account for the trading fees.

News information and stock prices are high–frequency data. Recent research has used ANNs or HMMs to analyse such high–frequency financial data. For instance, Lahmiri (2014) presents a forecasting model that integrates the discrete wavelet transforming and backpropagation neural networks for predicting financial time series. The model uses low– and high–frequency components, obtained through the decomposition of the financial time series data by discrete wavelet transformation, as input variables for forecast future stock prices. Arévalo et al. (2016) use deep neural networks to forecast the next one–minute average price. The deep neural networks model is trained on the current time (hour and minute) as well as the n–lagged one–minute

pseudo–returns, price standard deviations and trend indicators. Cartea and Jaimungal (2013) employ an HMM to examine how the intraday dynamics of the stock market have changed and how to use this information to develop trading strategies at high frequencies.

These existing studies use high–frequency financial data, only considering stock prices to predict stock price movement. In our future work, we will consider high–frequency news information and stock prices to build a prediction model based on neural networks or HMMs.

Appendix: RavenPack News Analytics (RPNA)

In this research, we use data on public news from RPNA, Edition 3.0. RavenPack (http://www.ravenpack.com/) is one of the most well–known providers of news analytics data. Another well–known provider is Thomson Reuters News Analytics (TRNA). News analytics is a relatively new tool based on AI and designed to improve the understanding of news events.

RPNA collect corporate news items (from the year 2000) from all public sources, including the Dow Jones, *Barron's* and the *Wall Street Journal*. There are two categories of editions: Global Equities and Global Macro. Global Equities editions contain only entities that are classified as companies. These companies include over 40,000 listed stocks from the world's equity markets, which spread across the Americas, Europe and the Asia–Pacific. Global Macro editions contain all entities that are not classified as companies. RavenPack analyses news on over 200 economies, delivering data on more than 138,000 places, 2,500 financially relevant organisations, 155 currencies and 82 commodities. In this thesis, we use the Global Equities editions dataset.

There are 39 fields used to describe each news item in the Global Equities editions (and 27 fields per news item in the Global Macro editions); these include the time stamp, the entity ID, the company name, the relevance of the news, the event category, the event sentiment, the novelty of the news, the composite sentiment score of the news, the story event count and the story ID. Some of these fields, such as event category and event sentiment, are generated by RPNA through content analysis.

The ‘time stamp’ is used to record the UTC at which the news occurred. RavenPack uses the ‘relevance’ variable to differentiate between news items where the corporation is the main object of the original news source and news items where the name of the corporation is mentioned only tangentially. The relevance variable attributes values from 0 to 100 and a score of 100 indicates that an article is highly relevant.

For a news story with a relevance score of 100, the ENS represents its novelty value. The first story to report a categorised event will receive a novelty score of 100. The novelty scores of subsequent stories about the same event will follow a decay function (i.e., 100, 75, 56, 42, 32, 24, 18, 13, 10, 8, 6, 4, 3, 2, 2, 1, 1, 1, 1, 0 …).

The ESS measures whether a particular news item contains favourable or unfavourable information about the underlying corporation. This variable represents the news sentiment for a given entity, ranging from 0 to 100, where 0 indicates extremely negative news, 50 indicates neutral news and 100 indicates extremely positive news.

In terms of the sentiment, RavenPack uses a proprietary computational linguistic analysis algorithm to quantify positive and negative perceptions of facts and opinions reported in the textual content of the news (Shi et al., 2016a). The core of this algorithm can be divided into two steps. First, a group of financial experts manually tag a set of stories and build up a historical database of words, phrases, combinations and other word–level definitions that have affected the target company, market or asset class. Then, the text in the specific news story is compared with the historical database and the

sentiments score is generated by automated computer classification using a Bayes Classifier.

Many researchers use RPNA to analyse the financial market. For instance, Smales (2014) examines the market reaction of leading Australian stocks to stock–specific news flow over an extended period. The study concurs with previous literature that news items are critically relevant to identifying significant effects. Akbas et al. (2016) find that high short interest is predictive of negative public news based on RPNA. Shi et al. (2016a) analyse how the hourly return volatility of S&P100 stocks from 2000 to 2010 are linked to the various linguistics–based sentiment scores of the news releases.

Bibliography

ABU-MOSTAFA, Y. S. & ATIYA, A. F. 1996. Introduction to financial forecasting. *Applied Intelligence,* 6**,** 205-213.

AGA, M. & KOCAMAN, B. 2008. Efficient market hypothesis and emerging capital markets: empirical evidence from Istanbul stock exchange. *International Research Journal of Finance and Economics,* 13**,** 131-144.

AGHABOZORGI, S. & TEH, Y. W. 2014. Stock market co-movement assessment using a three-phase clustering method. *Expert Systems with Applications,* 41**,** 1301-1314.

AKBAS, F., BOEHMER, E., ERTURK, B. & SORESCU, S. 2016. Short Interest, Returns, and Unfavorable Fundamental Information. *Financial Management*.

AKTAS, N., DE BODT, E. & VAN OPPENS, H. 2008. Legal insider trading and market efficiency. *Journal of Banking & Finance,* 32**,** 1379-1392.

AL-ANASWAH, N. & WILFLING, B. 2011. Identification of speculative bubbles using state-space models with Markov-switching. *Journal of Banking & Finance,* 35**,** 1073-1086.

ALAM, M. M. & UDDIN, M. G. S. 2009. Relationship between interest rate and stock price: empirical evidence from developed and developing countries. *International journal of business and management,* 4**,** 43.

ALANYALI, M., MOAT, H. S. & PREIS, T. 2013. Quantifying the relationship between financial news and the stock market. *Sci. Rep,* 3**,** 3578.

ALIZADEH, A. & NOMIKOS, N. 2004. A Markov regime switching approach for hedging stock indices. *Journal of Futures Markets,* 24**,** 649-674.

ALLEN, D. E., MCALEER, M. J. & SINGH, A. K. 2015. Machine news and volatility: the Dow Jones industrial average and the TRNA real-time high-frequency sentiment series. *The Handbook of High Frequency Trading***,** 327-344.

AMIHUD, Y. & WOHL, A. 2004. Political news and stock prices: The case of Saddam Hussein contracts. *Journal of banking & Finance,* 28**,** 1185-1200.

ANDREÃO, R. V., DORIZZI, B. & BOUDY, J. 2006. ECG signal analysis through hidden Markov models. *IEEE Transactions on Biomedical engineering,* 53**,** 1541-1549.

ANG, A. & BEKAERT, G. 2002. Regime switches in interest rates. *Journal of Business & Economic Statistics,* 20**,** 163-182.

ANN, W. K., SEQUEIRA, J. M. & MCALEER, M. 2005. Modelling the information content in insider trades in the Singapore exchange. *Mathematics and*

Computers in Simulation, 68**,** 417-428.

ANTWEILER, W. & FRANK, M. Z. 2004. Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance,* 59**,** 1259-1294.

ARÉVALO, A., NIÑO, J., HERNÁNDEZ, G. & SANDOVAL, J. High-frequency trading strategy based on deep neural networks. International conference on intelligent computing, 2016. Springer, 424-436.

ASAFU-ADJAYE, J. 2000. The relationship between energy consumption, energy prices and economic growth: time series evidence from Asian developing countries. *Energy economics,* 22**,** 615-625.

ASGHARIAN, H., HOLMFELDT, M. & LARSON, M. 2011. An event study of price movements following realized jumps. *Quantitative Finance,* 11**,** 933-946.

ATSALAKIS, G. S. & VALAVANIS, K. P. 2009. Surveying stock market forecasting techniques–Part II: Soft computing methods. *Expert Systems with Applications,* 36**,** 5932-5941.

AWOKUSE, T. O. 2007. Causality between exports, imports, and economic growth: Evidence from transition economies. *Economics Letters,* 94**,** 389-395.

BAHRAMMIRZAEE, A. 2010. A comparative survey of artificial intelligence applications in finance: artificial neural networks, expert system and hybrid intelligent systems. *Neural Computing and Applications,* 19**,** 1165-1195.

BALKE, N. S. & WOHAR, M. E. 2002. Low-frequency movements in stock prices: A state-space decomposition. *Review of Economics and Statistics,* 84**,** 649-667.

BARBERIS, N., SHLEIFER, A. & VISHNY, R. 1998. A model of investor sentiment. *Journal of financial economics,* 49**,** 307-343.

BEALE, M. H., HAGAN, M. T. & DEMUTH, H. B. 2015. Neural network toolbox user's guide.

BENEISH, M. D. & GARDNER, J. C. 1995. Information Costs and Liquidity Effects from Changes in the Dow Jones Industrial Average List. *Journal of Financial and Quantitative Analysis,* 30**,** 135-157.

BERA, A. K. & PARK, S. Y. 2008. Optimal portfolio diversification using the maximum entropy principle. *Econometric Reviews,* 27**,** 484-512.

BESSEMBINDER, H. & CHAN, K. 1995. The profitability of technical trading rules in the Asian stock markets. *Pacific-Basin Finance Journal,* 3**,** 257-284.

BETTMAN, J., HALLETT, A. & SAULT, S. 2010. Rumortrage: Can Investors Profit on Takeover Rumors on Internet Stock Message Boards? *Social Science Research Network Working Paper Series*.

BEYHAGHI, M. & HAWLEY, J. P. 2013. Modern portfolio theory and risk

management: assumptions and unintended consequences. *Journal of Sustainable Finance & Investment,* 3**,** 17-37.

BILMES, J. A. 1998. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *International Computer Science Institute,* 4**,** 126.

BOLLEN, J., MAO, H. & ZENG, X. 2011. Twitter mood predicts the stock market. *Journal of Computational Science,* 2**,** 1-8.

BOLLEN, N. P. 1998. Valuing options in regime-switching models. *The Journal of Derivatives,* 6**,** 38-49.

BOLLEN, N. P., GRAY, S. F. & WHALEY, R. E. 2000. Regime switching in foreign exchange rates:: Evidence from currency option prices. *Journal of Econometrics,* 94**,** 239-276.

BORGES, M. R. 2010. Efficient market hypothesis in European stock markets. *The European Journal of Finance,* 16**,** 711-726.

BOUBAKER, S., FARAG, H. & NGUYEN, D. K. 2015. Short-term overreaction to specific events: evidence from an emerging market. *Research in international business and finance,* 35**,** 153-165.

BRAV, A., GECZY, C. & GOMPERS, P. A. 2000. Is the abnormal return following equity issuances anomalous? *Journal of Financial Economics,* 56**,** 209-249.

BRIEC, W., KERSTENS, K. & JOKUNG, O. 2007. Mean-variance-skewness portfolio performance gauging: a general shortage function and dual approach. *Management science,* 53**,** 135-149.

BROTHERSON, W. T., EADES, K. M., HARRIS, R. S. & HIGGINS, R. C. 2015. 'Best Practices' in Estimating the Cost of Capital: An Update. *Journal of Applied Finance,* 23**,** 19.

BUFFINGTON, J. & ELLIOTT, R. J. 2002. American options with regime switching. *International Journal of Theoretical and Applied Finance,* 5**,** 497-514.

CALAFIORE, G. C. 2008. Multi-period portfolio optimization with linear control policies. *Automatica,* 44**,** 2463-2473.

CAO, Q., LEGGIO, K. B. & SCHNIEDERJANS, M. J. 2005. A comparison between Fama and French's model and artificial neural networks in predicting the Chinese stock market. *Computers & Operations Research,* 32**,** 2499-2512.

CARTEA, Á. & JAIMUNGAL, S. 2013. Modelling asset prices for algorithmic and high-frequency trading. *Applied Mathematical Finance,* 20**,** 512-547.

CELIKYURT, U. & ÖZEKICI, S. 2007. Multiperiod portfolio optimization models in stochastic markets using the mean–variance approach. *European Journal of Operational Research,* 179**,** 186-202.

CERVELLÓ-ROYO, R., GUIJARRO, F. & MICHNIUK, K. 2015. Stock market trading rule based on pattern recognition and technical analysis: Forecasting the DJIA index with intraday data. *Expert Systems with Applications,* 42**,** 5963-5975.

CHAN, K. C., GUP, B. E. & PAN, M.-S. 1997. International Stock Market Efficiency and Integration: A Study of Eighteen Nations. *Journal of Business Finance & Accounting,* 24**,** 803-813.

CHAN, W. S. 2003. Stock price reaction to news and no-news: drift and reversal after headlines. *Journal of Financial Economics,* 70**,** 223-260.

CHANG, P.-C., WANG, D.-D. & ZHOU, C.-L. 2012. A novel model by evolving partially connected neural network for stock price trend forecasting. *Expert Systems with Applications,* 39**,** 611-620.

CHANG, T.-S. 2011. A comparative study of artificial neural networks, and decision trees for digital game content stocks price prediction. *Expert Systems with Applications,* 38**,** 14846-14851.

CHARLES, A. & DARNÉ, O. 2014. Large shocks in the volatility of the Dow Jones Industrial Average index: 1928–2013. *Journal of Banking & Finance,* 43**,** 188-199.

CHAWLA, G. K. 2014. Estimating Cost of Capital in Today's Economic Environment. *Journal of Business and Behavior Sciences,* 26**,** 102.

CHEN, L., DA, Z. & ZHAO, X. 2013. What Drives Stock Price Movements? *Review of Financial Studies,* 26**,** 841-876.

CHEN, M., MAO, S. & LIU, Y. 2014. Big Data: A Survey. *Mobile Networks and Applications,* 19**,** 171-209.

CHIOU-WEI, S. Z., CHEN, C.-F. & ZHU, Z. 2008. Economic growth and energy consumption revisited—evidence from linear and nonlinear Granger causality. *Energy Economics,* 30**,** 3063-3076.

CHRISTIE, A. A. 1982. The stochastic behavior of common stock variances: Value, leverage and interest rate effects. *Journal of financial Economics,* 10**,** 407-432.

CHUNG, S.-L., HUNG, C.-H. & YEH, C.-Y. 2012. When does investor sentiment predict stock returns? *Journal of Empirical Finance,* 19**,** 217-240.

CLEMENTS, M. P. & KROLZIG, H. M. 1998. A Comparison of the Forecast Performance of Markov ‐ switching and Threshold Autoregressive Models of US GNP. *The Econometrics Journal,* 1**,** 47-75.

COHEN, G. & CABIRI, E. 2015. Can technical oscillators outperform the buy and hold strategy? *Applied Economics,* 47**,** 3189-3197.

CREAMER, G. G. 2015. Can a corporate network and news sentiment improve portfolio optimization using the Black–Litterman model? *Quantitative Finance,*

15**,** 1405-1416.

CUTLER, D. M., POTERBA, J. M. & SUMMERS, L. H. 1989. What moves stock prices? *The Journal of Portfolio Management,* 15**,** 4-12.

DACCO, R. & SATCHELL, S. 1999. Why do regime ‐ switching models forecast so badly? *Journal of Forecasting,* 18**,** 1-16.

DAHLQUIST, M. & GRAY, S. F. 2000. Regime-switching and interest rates in the European monetary system. *Journal of International Economics,* 50**,** 399-419.

DE OLIVEIRA, F. A., NOBRE, C. N. & ZÁRATE, L. E. 2013. Applying Artificial Neural Networks to prediction of stock price and improvement of the directional prediction index–Case study of PETR4, Petrobras, Brazil. *Expert Systems with Applications,* 40**,** 7596-7606.

DEMIGUEL, V., GARLAPPI, L. & UPPAL, R. 2009. Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy? *Review of Financial Studies,* 22**,** 1915-1953.

DEMPSTER, A. P., LAIRD, N. M. & RUBIN, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)***,** 1-38.

DICKEY, D. A. & FULLER, W. A. 1979. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association,* 74**,** 427-431.

DUEKER, M. J. 1997. Markov switching in GARCH processes and mean-reverting stock-market volatility. *Journal of Business & Economic Statistics,* 15**,** 26-34.

DURLAND, J. M. & MCCURDY, T. H. 1994. Duration-dependent transitions in a Markov model of US GNP growth. *Journal of Business & Economic Statistics,* 12**,** 279-288.

ELLIOTT, R. J., SIU, T. K. & BADESCU, A. 2010. On mean-variance portfolio selection under a hidden Markovian regime-switching model. *Economic modelling,* 27**,** 678-686.

ELTON, E. J. & GRUBER, M. J. 1997. Modern portfolio theory, 1950 to date. *Journal of Banking & Finance,* 21**,** 1743-1759.

ENGEL, C. 1994. Can the Markov switching model forecast exchange rates? *Journal of International Economics,* 36**,** 151-165.

ESTES, J. E., SAILER, C. & TINNEY, L. R. 1986. Applications of artificial intelligence techniques to remote sensing. *The Professional Geographer,* 38**,** 133-141.

FAMA, E. F. 1965. Random walks in stock market prices. *Financial Analysts Journal,* 21**,** 55-59.

FAMA, E. F. 1970. Efficient capital markets: A review of theory and empirical work. *Journal of Finance,* 25**,** 383-417.

FAMA, E. F. 1991. Efficient capital markets: II. *The journal of finance,* 46**,** 1575-1617.

FAN, K., SHEN, Y., SIU, T. K. & WANG, R. 2016. Pricing dynamic fund protection under hidden Markov models. *IMA Journal of Management Mathematics***,** dpw014.

FERNANDEZ-RODRıGUEZ, F., GONZALEZ-MARTEL, C. & SOSVILLA-RIVERO, S. 2000. On the profitability of technical trading rules based on artificial neural networks:: Evidence from the Madrid stock market. *Economics letters,* 69**,** 89-94.

FIESCHI, M. 2013. *Artificial intelligence in medicine: Expert systems,* Springer.

FLANNERY, M. J. & JAMES, C. M. 1984. The effect of interest rate changes on the common stock returns of financial institutions. *The Journal of Finance,* 39**,** 1141-1153.

FRANKS, J. R. & HARRIS, R. S. 1989. Shareholder Wealth Effects of Corporate Takeovers: The U.K. Experience 1955-1985. *Journal of Financial Economics,* 23 (2)**,** 225-249

GILLAM, R. A., GUERARD, J. B. & CAHAN, R. 2015. News volume information: Beyond earnings forecasting in a global stock selection model. *International Journal of Forecasting,* 31**,** 575-581.

GÖÇKEN, M., ÖZÇALıCı, M., BORU, A. & DOSDOĞRU, A. T. 2016. Integrating metaheuristics and artificial neural networks for improved stock price prediction. *Expert Systems with Applications,* 44**,** 320-331.

GOOGLE INC. 2015. Investor Relations.

GOZBASI, O., KUCUKKAPLAN, I. & NAZLIOGLU, S. 2014. Re-examining the Turkish stock market efficiency: Evidence from nonlinear unit root tests. *Economic Modelling,* 38**,** 381-384.

GRANGER, C. W. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica,* 37**,** 424-438.

GRANGER, C. W. 1988. Some recent development in a concept of causality. *Journal of econometrics,* 39**,** 199-211.

GRANGER, C. W., HUANGB, B.-N. & YANG, C.-W. 2000. A bivariate causality between stock prices and exchange rates: evidence from recent Asianflu. *The Quarterly Review of Economics and Finance,* 40**,** 337-354.

GRAY, S. F. 1996. Modeling the conditional distribution of interest rates as a regime-switching process. *Journal of Financial Economics,* 42**,** 27-62.

GÜLPıNAR, N. & RUSTEM, B. 2007. Worst-case robust decisions for multi-period

mean–variance portfolio optimization. *European Journal of Operational Research,* 183**,** 981-1000.

GUPTA, A. Stock market prediction using Hidden Markov Models. 2012 Students Conference on Engineering and Systems (SCES), 16-18 March 2012 2012.

GURESEN, E., KAYAKUTLU, G. & DAIM, T. U. 2011. Using artificial neural network models in stock market index prediction. *Expert Systems with Applications,* 38**,** 10389-10397.

HAAPANEN, R. & EK, A. R. 2001. Software and instructions for kNN applications in forest resources description and estimation.

HAERI, A., HATEFI, S. M. & REZAIE, K. 2015. Forecasting about EURJPY exchange rate using hidden Markova model and CART classification algorithm. *Journal of Advanced Computer Science & Technology,* 4**,** 84.

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

HALDRUP, N. & NIELSEN, M. Ø. 2006. A regime switching long memory model for electricity prices. *Journal of econometrics,* 135**,** 349-376.

HAMILTON, J. D. 1989. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the Econometric Society***,** 357-384.

HAN, E.-H. S., KARYPIS, G. & KUMAR, V. Text categorization using weight adjusted k-nearest neighbor classification. Pacific-asia conference on knowledge discovery and data mining, 2001. Springer, 53-65.

HAN, Y., YANG, K. & ZHOU, G. 2013. A new anomaly: The cross-sectional profitability of technical analysis. *Journal of Financial and Quantitative Analysis,* 48**,** 1433-1461.

HARRIS, L. 1991. Stock price clustering and discreteness. *Review of financial studies,* 4**,** 389-415.

HASSAN, M. R. & NATH, B. Stock market forecasting using hidden Markov model: a new approach. 5th International Conference on Intelligent Systems Design and Applications (ISDA'05), 2005. IEEE, 192-196.

HASSAN, M. R., NATH, B. & KIRLEY, M. 2007. A fusion model of HMM, ANN and GA for stock market forecasting. *Expert Systems with Applications,* 33**,** 171-180.

HENRIKSEN, P. N. 2011. Pricing barrier options by a regime switching model. *Quantitative Finance,* 11**,** 1221-1231.

HESTON, S. L. & SINHA, N. R. 2017. News vs. Sentiment: Predicting Stock Returns from News Stories. *Financial Analysts Journal,* 73**,** 1-17.

HIEMSTRA, C. & JONES, J. D. 1994. Testing for Linear and Nonlinear Granger Causality in the Stock Price - Volume Relation. *The Journal of Finance,* 49**,** 1639-1664.

HILL, T., MARQUEZ, L., O'CONNOR, M. & REMUS, W. 1994. Artificial neural network models for forecasting and decision making. *International Journal of Forecasting,* 10**,** 5-15.

HO, K.-Y., SHI, Y. & ZHANG, Z. 2013. How does news sentiment impact asset volatility? Evidence from long memory and regime-switching approaches. *The North American Journal of Economics and Finance,* 26**,** 436-456.

HO, K.-Y., WANG, K. & WANG, W. W. 2015. Foreign Trade and Economic Growth in Recent China: A Granger Causality Analysis. *A New Paradigm for International Business.* Springer.

HO, K.-Y. & WANG, W. W. 2016. Predicting Stock Price Movements with News Sentiment: An Artificial Neural Network Approach. *Artificial Neural Network Modelling.* Springer.

HOFFMANN, R., LEE, C. G., RAMASAMY, B. & YEUNG, M. 2005. FDI and pollution: a granger causality test using panel data. *Journal of international development,* 17**,** 311-317.

HOGAN, S., JARROW, R., TEO, M. & WARACHKA, M. 2004. Testing market efficiency using statistical arbitrage with applications to momentum and value strategies. *Journal of Financial economics,* 73**,** 525-565.

HSU, P.-H. & KUAN, C.-M. 2005. Reexamining the profitability of technical analysis with data snooping checks. *Journal of Financial Econometrics,* 3**,** 606-628.

HUA, Y. & WANG, X. 2014. Portfolio Selection with a Hidden Markov Model. *Quality Technology & Quantitative Management,* 11**,** 167-174.

HUANG, W., NAKAMORI, Y. & WANG, S.-Y. 2005. Forecasting stock market movement direction with support vector machine. *Computers & Operations Research,* 32**,** 2513-2522.

HUANG, X. D., ARIKI, Y. & JACK, M. A. 1990. *Hidden Markov models for speech recognition,* Edinburgh university press Edinburgh.

HUANG, Y., ZHOU, S., HUANG, K. & GUAN, J. Boosting financial trend prediction with twitter mood based on selective hidden Markov models. International Conference on Database Systems for Advanced Applications, 2015. Springer, 435-451.

HUI, E. C. & CHAN, K. K. K. 2014. Can we still beat "buy-and-hold" for individual stocks? *Physica A: Statistical Mechanics and its Applications,* 410**,** 513-534.

HUI, S. C. & JHA, G. 2000. Data mining for customer service support. *Information & Management,* 38**,** 1-13.

IBRAHIM, M. 1999. Macroeconomic variables and stock prices in Malaysia: an empirical analysis. *Asian Economic Journal,* 13**,** 219-231.

INGRAND, F. & GHALLAB, M. 2014. Robotics and artificial intelligence: A perspective on deliberation functions. *AI Communications,* 27**,** 63-80.

JENNINGS, R. & STARKS, L. 1986. Earnings announcements, stock price adjustment, and the existence of option markets. *The journal of Finance,* 41**,** 107-125.

JENSEN, M. C. 1969. Risk, the pricing of capital assets, and the evaluation of investment portfolios. *The Journal of Business,* 42**,** 167-247.

JENSEN, N. M. & SCHMITH, S. 2005. Market responses to politics: The rise of Lula and the decline of the Brazilian stock market. *Comparative Political Studies,* 38**,** 1245-1270.

JOBSON, J. D. & KORKIE, B. 1980. Estimation for Markowitz efficient portfolios. *Journal of the American Statistical Association,* 75**,** 544-554.

JOERDING, W. 1986. Economic growth and defense spending: Granger causality. *Journal of Development Economics,* 21**,** 35-40.

JONES, D., CO & JONES-IRWIN, D. 1989. *The Dow Jones Investor's Handbook,* Dow Jones Books.

JOULIN, A., LEFEVRE, A., GRUNBERG, D. & BOUCHAUD, J. P. 2008. Stock price jumps: news and volume play a minor role. *Arxiv Preprint Arxiv:0803.1769*.

KARA, Y., BOYACIOGLU, M. A. & BAYKAN, Ö. 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**,** 5311-5319.

KIM, H. & MEI, J. 2001. What makes the stock market jump? An analysis of political risk on Hong Kong stock returns. *Journal of International Money and Finance,* 20**,** 1003-1016.

KIM, K.-J. & HAN, I. 2000. Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. *Expert systems with Applications,* 19**,** 125-132.

KONNO, H., SHIRAKAWA, H. & YAMAZAKI, H. 1993. A mean-absolute deviation-skewness portfolio optimization model. *Annals of Operations Research,* 45**,** 205-220.

KONNO, H. & YAMAZAKI, H. 1991. Mean-absolute deviation portfolio optimization model and its applications to Tokyo stock market. *Management science,* 37**,** 519-531.

KONONENKO, I. 2001. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine,* 23**,** 89-109.

KORN, R. & KORN, E. 2001. *Option pricing and portfolio optimization: modern methods of financial mathematics*, American Mathematical Soc.

KOSKI, A. 1996. Modelling ECG signals with hidden Markov models. *Artificial intelligence in medicine,* 8**,** 453-471.

KOUTMOS, G. 1996. Modeling the dynamic interdependence of major European stock markets. *Journal of Business Finance & Accounting,* 23**,** 975-988.

KRISTOUFEK, L. 2013. Can Google Trends search queries contribute to risk diversification? *Scientific reports,* 3**,** 5.

KUMAR, M., JINDAL, M. & SHARMA, R. k-nearest neighbor based offline handwritten Gurmukhi character recognition. Image Information Processing (ICIIP), 2011 International Conference on, 2011. IEEE, 1-4.

KYLE, A. S. 1985. Continuous auctions and insider trading. *Econometrica,* 53**,** 1315-1335.

LAHMIRI, S. 2014. Wavelet low-and high-frequency components as features for predicting stock prices with backpropagation neural networks. *Journal of King Saud University-Computer and Information Sciences,* 26**,** 218-227.

LAI, R. K., FAN, C.-Y., HUANG, W.-H. & CHANG, P.-C. 2009. Evolving and clustering fuzzy decision tree for financial time series data forecasting. *Expert Systems with Applications,* 36**,** 3761-3773.

LAI, T. L., XING, H. & CHEN, Z. 2011. Mean-variance portfolio optimization when means and covariances are unknown. *The Annals of Applied Statistics***,** 798-823.

LAM, P. S. 2004. A Markov ‐ Switching Model Of Gnp Growth With Duration Dependence. *International Economic Review,* 45**,** 175-204.

LANDSMAN, Z. 2010. On the tail mean–variance optimal portfolio selection. *Insurance: Mathematics and Economics,* 46**,** 547-553.

LARY, D. J., ALAVI, A. H., GANDOMI, A. H. & WALKER, A. L. 2016. Machine learning in geosciences and remote sensing. *Geoscience Frontiers,* 7**,** 3-10.

LEE, C.-C., LEE, J.-D. & LEE, C.-C. 2010. Stock prices and the efficient market hypothesis: Evidence from a panel stationary test with structural breaks. *Japan and the World Economy,* 22**,** 10.

LEE, C. F. 1977. Functional form, skewness effect, and the risk-return relationship. *Journal of financial and quantitative analysis,* 12**,** 55-72.

LEE, C. G. 2009. Foreign direct investment, pollution and economic growth: evidence from Malaysia. *Applied Economics,* 41**,** 1709-1716.

LEE, Y., OW, L. T. C. & LING, D. N. C. Hidden markov models for forex trends prediction. Information Science and Applications (ICISA), 2014

International Conference on, 2014. IEEE, 1-4.

LEIGH, W., FROHLICH, C. J., HORNIK, S., PURVIS, R. L. & ROBERTS, T. L. 2008. Trading with a stock chart heuristic. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans,* 38**,** 93-104.

LEIGH, W., PURVIS, R. & RAGUSA, J. M. 2002. Forecasting the NYSE composite index with technical analysis, pattern recognizer, neural network, and genetic algorithm: a case study in romantic decision support. *Decision support systems,* 32**,** 361-377.

LI, Q., WANG, T., LI, P., LIU, L., GONG, Q. & CHEN, Y. 2014. The effect of news and public mood on stock movements. *Information Sciences,* 278**,** 826-840.

LI, X., XIE, H., WANG, R., CAI, Y., CAO, J., WANG, F., MIN, H. & DENG, X. 2016. Empirical analysis: stock market prediction via extreme learning machine. *Neural Computing and Applications,* 27**,** 67-78.

LIAO, S.-H., CHU, P.-H. & HSIAO, P.-Y. 2012. Data mining techniques and applications–A decade review from 2000 to 2011. *Expert Systems with Applications,* 39**,** 11303-11311.

LIM, K. P. & BROOKS, R. 2011. The evolution of stock market efficiency over time: a survey of the empirical literature. *Journal of Economic Surveys,* 25**,** 69-108.

LIU, L., WU, J., LI, P. & LI, Q. 2015. A social-media-based approach to predicting stock comovement. *Expert Systems with Applications,* 42**,** 3893-3901.

LUCAS, D. J. & MCDONALD, R. L. 1990. Equity issues and stock price dynamics. *The journal of finance,* 45**,** 1019-1043.

LUKASHIN, A. V. & BORODOVSKY, M. 1998. GeneMark. hmm: new solutions for gene finding. *Nucleic acids research,* 26**,** 1107-1115.

LUND, H. H. Adaptive robotics in the entertainment industry. Computational Intelligence in Robotics and Automation, 2003. Proceedings. 2003 IEEE International Symposium on, 2003. IEEE, 595-602.

LUX, T. 2011. Sentiment dynamics and stock returns: the case of the German stock market. *Empirical economics,* 41**,** 663-679.

MALATESTA, P. H. & THOMPSON, R. 1985. Partially anticipated events: A model of stock price reactions with an application to corporate acquisitions. *Journal of Financial Economics,* 14**,** 237-250.

MANYIKA, J., CHUI, M., BROWN, B., BUGHIN, J., DOBBS, R., ROXBURGH, C. & BYERS, A. H. 2011. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute.

MARKOWITZ, H. 1952. Portfolio selection. *The journal of finance,* 7**,** 77-91.

MARKOWITZ, H. M. 1959. *Portfolio selection: efficient diversification of*

investments, Yale university press.

MARSH, I. W. 2000. High ‐ frequency Markov switching models in the foreign exchange market. *Journal of Forecasting,* 19**,** 123-134.

MASIP, D. & VITRIÀ, J. 2008. Shared feature extraction for nearest neighbor face recognition. *IEEE Transactions on Neural Networks,* 19**,** 586-595.

MAYER-SCHÖNBERGER, V. & CUKIER, K. 2013. *Big data: A revolution that will transform how we live, work, and think*, Houghton Mifflin Harcourt.

MCADAM, P. & MCNELIS, P. 2005. Forecasting inflation with thick models and neural networks. *Economic Modelling,* 22**,** 848-867.

MILLER, A., BLOTT, B. & HAMES, T. 1992. Review of neural network applications in medical imaging and signal processing. *Medical and Biological Engineering and Computing,* 30**,** 449-464.

MITRA, G. & MITRA, L. 2011. *The handbook of news analytics in finance*, John Wiley & Sons.

MITRA, L., MITRA, G. & DIBARTOLOMEO¶, D. 2009. Equity portfolio risk estimation using market information and sentiment. *Quantitative Finance,* 9**,** 887-895.

MLAMBO, C. & BIEKPE, N. 2007. The efficient market hypothesis: Evidence from ten African stock markets. *Investment Analysts Journal,* 36**,** 5-17.

MOORE, T. & WANG, P. 2007. Volatility in stock returns for new EU member states: Markov regime switching model. *International Review of Financial Analysis,* 16**,** 282-292.

MOUNT, T. D., NING, Y. & CAI, X. 2006. Predicting price spikes in electricity markets using a regime-switching model with time-varying parameters. *Energy Economics,* 28**,** 62-80.

NARAYAN, P. K., NARAYAN, S. & THURAISAMY, K. S. 2014. Can institutions and macroeconomic factors predict stock returns in emerging markets? *Emerging Markets Review,* 19**,** 77-95.

NAZÁRIO, R. T. F., E SILVA, J. L., SOBREIRO, V. A. & KIMURA, H. 2017. A Literature Review Of Technical Analysis On Stock Markets. *The Quarterly Review of Economics and Finance*.

NEAL, R. & WHEATLEY, S. M. 1998. Do measures of investor sentiment predict returns? *Journal of Financial and Quantitative Analysis,* 33**,** 523-547.

NGUYEN-TUONG, D. & PETERS, J. 2011. Model learning for robot control: a survey. *Cognitive processing,* 12**,** 319-340.

NI, L.-P., NI, Z.-W. & GAO, Y.-Z. 2011. Stock trend prediction based on fractal feature selection and support vector machine. *Expert Systems with Applications,* 38**,**

5569-5576.

NISAR, S. & HANIF, M. 2012. Testing weak form of efficient market hypothesis: Empirical evidence from South-Asia. *World Applied Sciences Journal,* 17**,** 414-427.

NOFSINGER, J. R. 2005. Social mood and financial economics. *The Journal of Behavioral Finance,* 6**,** 144-160.

ÖZATAY, F., ÖZMEN, E. & ŞAHINBEYOĞLU, G. 2009. Emerging market sovereign spreads, global financial conditions and US macroeconomic news. *Economic Modelling,* 26**,** 526-531.

PALIWAL, M. & KUMAR, U. A. 2009. Neural networks and statistical techniques: A review of applications. *Expert systems with applications,* 36**,** 2-17.

PARANJAPE-VODITEL, P. & DESHPANDE, U. 2013. A stock market portfolio recommender system based on association rule mining. *Applied Soft Computing,* 13**,** 1055-1063.

PARK, C. H. & IRWIN, S. H. 2007. What do we know about the profitability of technical analysis? *Journal of Economic Surveys,* 21**,** 786-826.

PATEL, J., SHAH, S., THAKKAR, P. & KOTECHA, K. 2015. Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Systems with Applications,* 42**,** 259-268.

PETRUSHIN, V. A. Hidden markov models: Fundamentals and applications. Online Symposium for Electronics Engineer, 2000.

POUND, J. & ZECKHAUSER, R. 1990. Clearly heard on the street: The effect of takeover rumors on stock prices. *Journal of Business,* 63**,** 291-308.

PREETHI, G. & SANTHI, B. 2012. STOCK MARKET FORECASTING TECHNIQUES: A SURVEY. *Journal of Theoretical & Applied Information Technology,* 46.

QI, M. 1996. Financial applications of Artificial Neural Networks. *Handbook of Statistics,* 14**,** 529-552.

QUAH, T.-S. 2008. DJIA stock selection assisted by neural network. *Expert Systems with Applications,* 35**,** 50-58.

RABINER, L. & JUANG, B. 1986. An introduction to hidden Markov models. *ieee assp magazine,* 3**,** 4-16.

RABINER, L. R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE,* 77**,** 257-286.

RAMESH, A., KAMBHAMPATI, C., MONSON, J. & DREW, P. 2004. Artificial intelligence in medicine. *Annals of The Royal College of Surgeons of England,* 86**,** 334.

RAY, S. 2012. Testing Granger Causal Relationship between Macroeconomic Variables and Stock Price Behaviour: Evidence from India. *Advances in Applied Economics and Finance,* 3**,** 470-481.

REESE, H., NILSSON, M., SANDSTRÖM, P. & OLSSON, H. 2002. Applications using estimates of forest parameters derived from satellite and forest inventory data. *Computers and Electronics in Agriculture,* 37**,** 37-55.

RODRÍGUEZ, G., SORIA, Á. & CAMPO, M. 2016. Artificial intelligence in service-oriented software design. *Engineering Applications of Artificial Intelligence,* 53**,** 86-104.

ROGALSKI, R. J. 1978. The dependence of prices and volume. *The Review of Economics and Statistics***,** 268-274.

ROSS, S. A. 1976. The arbitrage theory of capital asset pricing. *Journal of economic theory,* 13**,** 341-360.

SCHERER, B. 2002. *Portfolio construction and risk budgeting,* Risk Books.

SCHMELING, M. 2009. Investor sentiment and stock returns: Some international evidence. *Journal of Empirical Finance,* 16**,** 394-408.

SCHWERT, G. W. 1989. Why does stock market volatility change over time? *The journal of finance,* 44**,** 1115-1153.

SHARPE, W. F. 1964. Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance,* 19**,** 425-442.

SHARPE, W. F. 1966. Mutual fund performance. *The Journal of business,* 39**,** 119-138.

SHEN, Y., FAN, K. & SIU, T. K. 2014. Option Valuation Under a Double Regime ‐ Switching Model. *Journal of Futures Markets,* 34**,** 451-478.

SHI, S. & SONG, Y. 2014. Identifying speculative bubbles using an infinite hidden Markov model. *Journal of Financial Econometrics***,** nbu025.

SHI, Y. & HO, K.-Y. 2015. Long memory and regime switching: A simulation study on the Markov regime-switching ARFIMA model. *Journal of Banking & Finance,* 61**,** S189-S204.

SHI, Y., HO, K.-Y. & LIU, W.-M. 2016a. Public information arrival and stock return volatility: Evidence from news sentiment and Markov Regime-Switching Approach. *International Review of Economics & Finance,* 42**,** 291-312.

SHI, Y., LIU, W.-M. & HO, K.-Y. 2016b. Public news arrival and the idiosyncratic volatility puzzle. *Journal of Empirical Finance,* 37**,** 159-172.

SICILIANO, B. & KHATIB, O. 2016. *Springer handbook of robotics,* Springer.

SKINNER, D. 1994. Why firms voluntarily disclose bad news. *Journal of Accounting*

Research, 32**,** 38-60.

ŠKOVRÁNEK, T., PODLUBNY, I. & PETRÁŠ, I. 2012. Modeling of the national economies in state-space: A fractional calculus approach. *Economic Modelling,* 29**,** 1322-1327.

SLOAN, R. 1996. Do stock prices fully reflect information in accruals and cash flows about future earnings?(Digest summary). *Accounting review,* 71**,** 289-315.

SMALES, L. A. 2014. Non-scheduled news arrival and high-frequency stock market dynamics: Evidence from the Australian Securities Exchange. *Research in International Business and Finance,* 32**,** 122-138.

SNELL, A. & TONKS, I. 1998. Testing for asymmetric information and inventory control effects in market maker behaviour on the London Stock Exchange. *Journal of Empirical Finance,* 5**,** 1-25.

SPECHT, D. F. Probabilistic neural networks for classification, mapping, or associative memory. IEEE international conference on neural networks, 1988. 525-532.

SPECHT, D. F. 1990. Probabilistic neural networks. *Neural networks,* 3**,** 109-118.

SU, D. 2003. Stock price reactions to earnings announcements: evidence from Chinese markets. *Review of Financial Economics,* 12**,** 271-286.

SZOLOVITS, P., PATIL, R. S. & SCHWARTZ, W. B. 1988. Artificial intelligence in medical diagnosis. *Annals of internal medicine,* 108**,** 80-87.

TAKANO, Y. & GOTOH, J.-Y. 2014. Multi-period portfolio selection using kernel-based control policy with dimensionality reduction. *Expert Systems with Applications,* 41**,** 3901-3914.

TAN, S. 2006. An effective refinement strategy for KNN text classifier. *Expert Systems with Applications,* 30**,** 290-298.

TANAKA-YAMAWAKI, M. & TOKUOKA, S. 2007. Adaptive use of technical indicators for the prediction of intra-day stock prices. *Physica A: Statistical Mechanics and its Applications,* 383**,** 125-133.

TAO, L., ELHAMIFAR, E., KHUDANPUR, S., HAGER, G. D. & VIDAL, R. Sparse hidden markov models for surgical gesture classification and skill evaluation. International Conference on Information Processing in Computer-Assisted Interventions, 2012. Springer, 167-177.

TAY, F. E. & CAO, L. 2001. Application of support vector machines in financial time series forecasting. *Omega,* 29**,** 309-317.

TAYLOR, S. J. 2011. *Asset price dynamics, volatility, and prediction,* Princeton university press.

TEIXEIRA, L. A. & DE OLIVEIRA, A. L. I. 2010. A method for automatic stock

trading combining technical analysis and nearest neighbor classification. *Expert Systems with Applications,* 37**,** 6885-6890.

TETLOCK, P. C. 2007. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance,* 62**,** 1139-1168.

TICKNOR, J. L. 2013. A Bayesian regularized artificial neural network for stock market forecasting. *Expert Systems with Applications,* 40**,** 5501-5506.

TIMMERMANN, A. 2012. Structural breaks, incomplete information, and stock prices. *Journal of Business & Economic Statistics*.

TREYNOR, J. L. 1965. How to rate management of investment funds. *Harvard business review,* 43**,** 63-75.

TSAI, I.-C. 2012. The relationship between stock price index and exchange rate in Asian markets: A quantile regression approach. *Journal of International Financial Markets, Institutions and Money,* 22**,** 609-621.

TUIA, D., MERENYI, E., JIA, X. & GRANA-ROMAY, M. 2014. Foreword to the special issue on machine learning for remote sensing data processing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* 7**,** 1007-1011.

TZIRALIS, G. & TATSIOPOULOS, I. 2012. Prediction markets: An extended literature review. *The journal of prediction markets,* 1**,** 75-91.

VAN NORDEN, S. & SCHALLER, H. 1999. Speculative behavior, regime-switching, and stock market crashes. *Nonlinear time series analysis of economic and financial data.* Springer.

VITERBI, A. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory,* 13**,** 260-269.

WANG, J.-L. & CHAN, S.-H. 2006. Stock market trading rule discovery using two-layer bias decision tree. *Expert Systems with Applications,* 30**,** 605-611.

WANG, J.-Z., WANG, J.-J., ZHANG, Z.-G. & GUO, S.-P. 2011. Forecasting stock indices with back propagation neural network. *Expert Systems with Applications,* 38**,** 14346-14355.

WANG, K., LI, M., HADLEY, D., LIU, R., GLESSNER, J., GRANT, S. F., HAKONARSON, H. & BUCAN, M. 2007. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome research,* 17**,** 1665-1674.

WANG, K. T. & WANG, W. W. 2017. Competition in the stock market with asymmetric information. *Economic Modelling,* 61**,** 40-49.

WANG, W. W., HO, K.-Y., LIU, W.-M. R. & WANG, K. T. The relation between news events and stock price jump: an analysis based on neural network. 20th

International Congress on Modelling and Simulation (MODSIM2013), 2013 Adelaide, Australia.

WESTERLUND, J. & NARAYAN, P. 2013. Testing the efficient market hypothesis in conditionally heteroskedastic futures markets. *Journal of Futures Markets,* 33**,** 1024-1045.

WESTERLUND, J., NORKUTE, M. & NARAYAN, P. K. 2015. A factor analytical approach to the efficient futures market hypothesis. *Journal of Futures Markets,* 35**,** 357-370.

WHITE, H. 1988. Economic prediction using neural networks: The case of IBM daily stock returns.

WONG, B. K., BODNOVICH, T. A. & SELVI, Y. 1997. Neural network applications in business: A review and analysis of the literature (1988–1995). *Decision Support Systems,* 19**,** 301-320.

WONG, B. K. & SELVI, Y. 1998. Neural network applications in finance: a review and analysis of literature (1990–1996). *Information & Management,* 34**,** 129-139.

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

YANG, S. Regression nearest neighbor in face recognition. Pattern Recognition, 2006. ICPR 2006. 18th International Conference on, 2006. IEEE, 515-518.

YEN, G. & LEE, C.-F. 2008. Efficient Market Hypothesis (EMH): Past, Present and Future. *Review of Pacific Basin Financial Markets and Policies,* 11**,** 305-329.

YI, K. & BEHESHTI, J. 2009. A hidden Markov model-based text classification of medical documents. *Journal of Information Science,* 35**,** 67-81.

YONG, Z., YOUWEN, L. & SHIXIONG, X. 2009. An improved KNN text classification algorithm based on clustering. *Journal of computers,* 4**,** 230-237.

YU, X.-P. & YU, X.-G. Novel text classification based on k-nearest neighbor. Machine Learning and Cybernetics, 2007 International Conference on, 2007. IEEE, 3425-3430.

ZANCHETTIN, C., BEZERRA, B. L. D. & AZEVEDO, W. W. A KNN-SVM hybrid model for cursive handwriting recognition. Neural Networks (IJCNN), The 2012 International Joint Conference on, 2012. IEEE, 1-8.

ZEINALIZADEH, N., SHOJAIE, A. A. & SHARIATMADARI, M. 2015. Modeling and analysis of bank customer satisfaction using neural networks approach. *International Journal of Bank Marketing,* 33**,** 717-732.

ZHANG, G., PATUWO, B. E. & HU, M. Y. 1998. Forecasting with artificial neural networks:: The state of the art. *International journal of forecasting,* 14**,** 35-62.

ZHOU, X. Y. & YIN, G. 2003. Markowitz's mean-variance portfolio selection with regime switching: A continuous-time model. *SIAM Journal on Control and Optimization,* 42**,** 1466-1482.

ZHU, X. & ZHU, J. 2013. Predicting stock returns: A regime-switching combination approach and economic links. *Journal of Banking & Finance,* 37**,** 4120-4133.

ZOUAOUI, M., NOUYRIGAT, G. & BEER, F. 2011. How Does Investor Sentiment Affect Stock Market Crises? Evidence from Panel Data. *Financial Review,* 46**,** 723-747.