Applied Natural Language Processing and Machine Learning in Algorithmic Trading

By: Rapheal Olaniyan

This dissertation is submitted for the degree of
Doctor of Philosophy

Supervisor: Dr Daniel Stamate

Data Science & Soft Computing Lab,
Computing Department,
Goldsmiths, University of London
London, UK

October 2021

Applied Natural Language Processing and Machine Learning in Algorithmic Trading

Abstract

The frequent ups and downs are characteristic of the stock market. The conventional predictive models that assume that investors act rationally have not been able to capture the irregularities in the stock market. They rely mainly on fundamental data such as assets, liabilities, among others. As such, these models seem to fail to capture the stock market trends that are extremely sensitive to social, economic, and political behavioural elements. As a result, behavioural finance is embraced to attempt to correct these model shortcomings by adding some factors to help capture the sentimental contagion which may be at play in determining the stock market. Many research works have attempted to establish this relationship between emotions and the stock market but, surprisingly, findings from these works have been rather conflicting due to the generic nature of sentimental information. This thesis is therefore relevant in that it helps to clarify on the relationship between sentiments and the stock market. First, this work explores different sources of data including pre-processed sentiments and sentiments extracted directly from raw financial news data based on a proposed novel BERT-based Natural Language Processing (NLP) algorithm. Also, most of the previous studies claiming that emotions have predictive value on the stock market do so by developing various machine learning predictive models, but do not validate their claims rigorously. Such findings may be clearly misleading. This

problem is addressed in this thesis with a focus on the relevance and appropriateness of model applicability and statistical validation. Finally, this thesis proposes an approach that incorporates our proposed NLP and stock market trading algorithms. The NLP algorithm automatically extracts the sentiment polarities from financial news and activates the proposed stock market trading algorithm to predict the directions of the stock market prices.

Acknowledgments

My special thanks to ...

Family

I dedicate my dissertation work to my family. You have always been there to support me in every way. A special feeling of gratitude to my late dad for being there all the time. I am where I am because of you and I am forever grateful. I would always miss you and your kind heart.

PhD Supervisor

Dr. Daniel Stamate, many thanks for your support, guidance and valuable feedback. You awakened the data science in me. I learned a lot from you and I am very grateful.

Data Science and Soft Computing Lab, London

I would like to thank the team of the lab, for the support I received over the years especially with research and computing facilities , and for collaborations in particular with Fred Marechal and research interns with the lab. .

Computing Department

I would like to thank the members of the Computing Department, in particular Dr. Ida Pu who was my second supervisor, and Dr. John Howroyd and Dr. Tony Russell-Rose who provided me with excellent feedback for my upgrade report. I

would like also to thank the IT team in particular Eamonn Martin who provided excellent support with the servers whenever I needed it for my research.

List of Publications

1. R. Olaniyan, D. Stamate and I. Pu, *A two-step optimised BERT-based NLP algorithm for extracting sentiment from financial news* , Proc. 17th International Conference on Artificial Intelligence Applications and Innovations (AIAI), Springer, accepted, to appear, 2021.

2. R. Olaniyan, D. Stamate, I. Pu, A. Zamyatin, A. Vashkel and F. Marechal, *Predicting S&P 500 based on its constituents and their social media derived sentiment* , Proc. 11th International Conference on Computational Collective Intelligence (ICCCI), 2019, Springer LNCS, DOI: https://doi.org/10.1007/978-3-030-28377-3_12.

3. F. Marechal, D. Stamate, R. Olaniyan and J. Marek, *On XLE index constituents' social media based sentiment informing the index trend and volatility prediction*. Proc. 10th International Conference on Computational Collective Intelligence (ICCCI), 2018, Springer LNCS, DOI: https://doi.org/10.1007/978-3-319-98446-9_34.

4. R. Olaniyan, D. Stamate,D. Logofatu and L. Ouarbya, *Sentiment and Stock Market Volatility Predictive Modelling - a Hybrid Approach*. Proceedings of the 2nd IEEE International Conference on Data Science and Advanced Analytics (IEEE DSAA), 2015, Publisher: IEEE, DOI: 10.1109/DSAA.2015.7344855.

5. F. Murtagh, R. Olaniyan and D. Stamate, *A Novel Statistical and Machine Learning Hybrid Approach to Predicting S&P 500 using Sentiment Analysis*. Proceedings of the 8th International Conference of the ERCIM Working Group on Computational and Methodological Statistics, and International Conference on Computational and Financial Econometrics, 2015.

6. R. Olaniyan, D. Stamate and D. Logofatu, *Social Web-Based Anxiety Index's Predictive Information on S&P 500, Revisited*. Proceedings of the 3rd International Symposium on Statistical Learning and Data Sciences (SLDS), 2015, Springer LNAI, DOI: https://doi.org/10.1007/978-3-319-17091-6_15.

Contents

1 Introduction **7**

1.1 Background and motivation 7

1.2 Research questions addressed in the thesis 12

1.2.1 Research objective and scope 12

1.2.2 Structure of the thesis and contributions 13

2 Social web-based anxiety index's predictive information on S&P 500 revisited **20**

2.1 Discussion on the Web blog based Anxiety Index 22

2.1.1 Findings and limitations 24

2.2 Anxiety Index's predictive information on the stock market, revisited 26

2.3 Conclusion 31

3 Sentiment and stock market volatility predictive modelling - a hybrid approach **34**

3.1 Stock market and sentiment 36

3.1.1 Conventional Granger causality between sentiment and stock returns 38

3.1.2 Extending the approach to non-parametric non-linear Granger causality 43

3.2 A hybrid approach to predicting stock volatility 44

3.2.1 Sensitivity analysis 50
3.3 Discussion and conclusion 54

4 Predicting S&P 500 based on its constituents and their social media derived sentiment **57**
4.1 Motivation 57
4.2 Introduction 58
4.3 Stock data and sentiment information 59
4.3.1 Reducing data dimensionality 60
4.4 Sentiment's predictive information on S&P 500 64
4.4.1 Granger causality test: the linear model 64
4.4.2 Granger causality test: the nonlinear model 66
4.5 Jordan and Elman neural network based approach to predicting S&P500 with sentiment 67
4.6 Evolutionary optimised trading model 69
4.7 Discussion and conclusion 72

5 A two-step optimised BERT-based NLP algorithm for extracting sentiment from financial news **74**
5.1 Motivation 74
5.2 Introduction 75
5.3 Proposed approach 78
5.4 BERT NLP 79
5.5 Methodology 80
5.6 Primary NLP model 83
5.7 Secondary NLP model 85
5.8 Discussion and conclusion 87

6 Event-based algorithmic intraday trading **90**
6.1 Motivation 90
6.2 Introduction 91
6.3 Methodology 94

6.3.1 Sampling methodology 95
6.3.2 Stock market variables 102
6.3.3 Technical analysis variables 102
6.4 Stationarity 107
6.4.1 Log differencing for stationarity 108
6.4.2 Proposed fractional differencing technique for stationarity 108
6.5 Stock market predictive models 110
6.6 Discussion and conclusion 115

7 Event-based algorithmic intraday trading with applied natural language processing algorithms **118**
7.1 Motivation 118
7.2 Introduction 119
7.3 Stock data and sentiment information 122
7.3.1 Sentiment variables 124
7.3.2 Stationarity 126
7.4 Model development and applications 127
7.5 Discussion and conclusion 129

8 Contracts for Difference (CFDs) machine learning algorithm for optimising portfolios **131**
8.1 Motivation 131
8.2 Introduction 132
8.3 Methodology 134
8.4 Model application and findings 138
8.5 Discussion and conclusion 141

9 Conclusion **143**
9.1 Summary of the thesis 144
9.2 Summary of contributions 148
9.3 Constraints and limitations 150
9.4 Future research directions 151

References 169

Appendix A Appendix **169**

1 Introduction

1.1 Background and motivation

Stock market prediction is a subject that earns attention from researchers, traders, stockbrokers, and policy makers, just to mention a few. Researchers would like to develop reliably working predictive models, traders and stockbrokers would like to use the developed models, and policy makers may be keen to understand how the models impact businesses and the economy. The interests of these parties are clearly diverse, but they all have something in common: to have a predictive model that is reliable. All starts with identifying the key stock market indicators and using these to predict the stock market trends and future prices with the aim of beating the market and earning profits in the highly volatile and complex stock market; hence, the need for the stock market predictive model development. Several such predictive modelling approaches have been developed under the

assumption that historical stock prices and volume data can be used to predict future stock prices. Deng and Sakurai [118] proposed a set of trading rules based on the model developed using purely the traditional stock data and volume. Their findings suggest that the data is sufficient in developing profitable trading rules. Neto et al. [97] developed a predictive model using the stock market data and artificial neural network (NN) algorithms. A series of NN models were compared by using the mean square error (MSE) [26], receiver operating characteristics (ROC) [126] and absolute mean square error. They reached the conclusion that with their models they were able to predict the stock market. Wang et al. [132] studied the composite price performance of 225 highly capitalized stocks trading on the Tokyo Stock Exchange (TSE). The study combined some macroeconomic variables such as GDP, interest rate, consumer price index, short-term interest rate, long-term interest rate, among others, with the stock market data. The predictive model explored in the research work made use of the support vector machine (SVM) ([138]) framework with the aim of predicting the future directions of the stock index. Findings from the work also support the sufficiency of the data in stock market modelling.

Most of these models are classified as standard finance models in view of their strong assumption that the past stock market information has correlation with the future prices, and hence, the directions of the future prices can be determined using solely the historical data.

But come to think of it, if the past stock data information is sufficient to capture the future directions of the stock market, and that there are many rational stock market players in the market, would there still be any possibility of making profits in the market? If all the players are expected to act rationally and employ predictive models, then the market would be expected to be operating under the Efficient Market Hypothesis (EMH) [37] with no profit. Clearly, these are models built with the assumption of investor rationality. But the long-held belief in such models is now being questioned: the models appear too basic judging by their inefficiency in capturing the complex and dynamic nature of the stock market as the stock market returns and investor behaviour diverge away from the

fundamental prices and rationality respectively. In fact, some renowned financial experts with years of experience in financial modelling strongly doubt the predictive power of models.
For example, UBS global chief economist, Paul Donovan [57], states point-blank:

> "Economists should not forecast.
>
> Economic models are not precise. Models use lots of assumptions. Those assumptions may not turn out to be true. Models give a range of possibilities rather than a single, certain number."

They consider most variables used in modelling as mere noise with no statistical significance. That is, all financial data is purely nonstationary and noisy as a result of the volatile nature of the stock market. These call for attention in behavioural finance ([99],[42]) to resolve the shortcomings of the standard finance models. Behavioural finance relaxes the assumption that investors act rationally. It underlines the importance of sentimental contagion in investment. Since then, researchers have been focusing on the relationship between sentiment and the stock market. Shiller [114] opposed the EMH by stating that factors related to the field of behavioural finance influence the stock market as a result of some psychological contagion which makes investors to overreact or underreact. Sprenger et al. [130] also disagreed with the EMH. They argue that the stock market is inefficient and therefore abnormal returns can be earned. They implied that investors have the tendency to underreact or overreact to new information, which could be in the form of business news, online social networking blogs, and other forms of online expressions.

Observations from related research works sprang up interest in advancing the standard finance models to include sentiment in the predictive model development with the aim of enhancing the model reliability and efficiency. Yet in order to statistically validate this inclusion, one needs to consider the source of the sentiment, examine its statistical significance and the Granger causality between the sentiment and the stock market variables by using appropriate

models ([31],[109]), and also the time sensitivity of the sentimental information in order to possibly benefit from the market imperfection [113]. Figure 1.1.1 emphasizes how the impact of relevantly sentimental information may disrupt the stock market. On the 1st of March, 2018, the former US president Donald Trump [51] announced that there would be tariff on steel import. After the announcement there was fear in the market that countries affected, especially China, would attempt to retaliate. This led to the downward movements in the volume and stock price of United States Steel Corp. This is clearly the influence of sentimental contagion on the stock market.

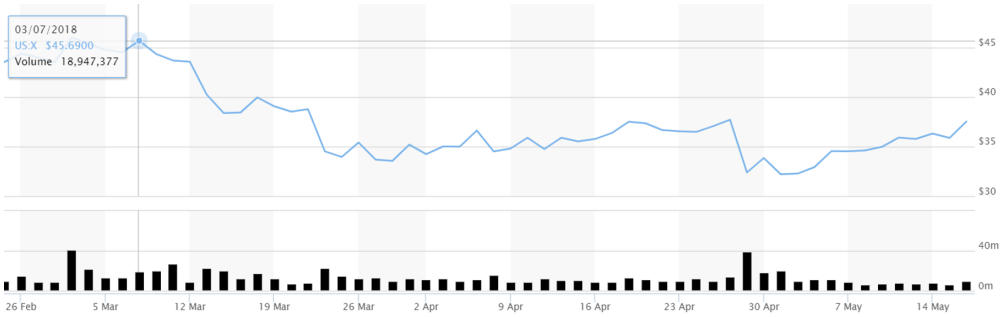

Figure 1.1.1: The former US president announced on the 1st of March, 2018, that he would impose tariff on the steel import. This shows the impact of the proposed tariff on United States Steel Corp. After the announcement there was downward movements in the stock price and stock volume. The picture is taken from [54].

In view of this observation, one may go in favour of the models that incorporate this element of information. But first, statistically investigating this hypothesis in order to establish its validation is very important. Luckily, a good number of studies have attempted to establish this relationship.

Oliviera et al. [103] examined if the sentiment from social blogs would statistically impact the stock market. In their study data from StockTwits, a microblogging platform that extracts sentiments from stock market-related blogs, was used as a proxy for sentiment. They analysed the impacts of the sentiment on the stock market returns, volatility ,and trading volume respectively. Their

findings show that the sentiment explored does not impact the stock market returns.

Smales [84] approached the study of the relationship between the financial news and the stock market in an interesting way. It measured the impacts of unscheduled financial news on the stock market using a linear framework with lengthy intraday financial data in the period 4th of January 2000 to 1st of November 2011. The paper concluded that sentiment from financial news has an influence on stock market activity, volatility, and spreads. Further disaggregating the financial news into negative and positive sentiments showed that the negative sentiment has greater impacts on the stock market than the positive sentiment. The intuition behind this finding is that investors respond more to negative sentiment than they do to positive sentiment since investment losses are more costly.

Bollen et al. [66] used the public mood conveyed from a large-scaled collection of tweets to measure the influence of sentiment on the stock market. In the process a Self-Organizing Fuzzy Neural Network model was employed and a Granger causality test was carried out. Their results supported the claim that emotions do influence the stock returns. However, Mao et al. [49] and Jahidul et al. [4] express some doubt about any possible non-linear Granger causality ([31],[109]) between the sentiment generated from the public mood and the stock market.

Gilbert and Kahahalios [38]'s work is one of the very few that have attempted to statistically validate their approach. It investigated the causal relationship between the stock market returns and sentiment and reached the same conclusion that sentiment influences the stock market returns. Results from [38], based on a collection of LiveJournal blogs, showed that sentiment possesses predictive information on the stock market returns.

As it stands, based on the aforementioned research work so far, one can hardly conclude on the statistical relevance of sentiment on the stock market - some work in favour and others against.

1.2 Research questions addressed in the thesis

So far, conclusions from related research works that explored the relationship between sentiments and the stock market have been clearly conflicting and unhelpful. Rather than solving the mystery behind this relationship, these works have exposed us to many questions that need answering:

1. Why do findings from many researches conflict each other in establishing the impacts of sentiment on the stock market?
2. Is it that the source of sentiment information explored could influence the relationship between sentiment and the stock market?
3. Could wrong model selection introduce a bias in examining this relationship leading to these conflicting findings?
4. Even with the right sentiment data source and appropriate model selection, could the established relationship be biased due to the time sensitivity of the sentiment information as strongly emphasized by the EMH?

These are therefore the research questions we attempt to answer in this thesis. Valid answers to the aforementioned questions can help to resolve these conflicting conclusions regarding the relevance of sentiment on the stock market.

1.2.1 Research objective and scope

Understanding the true impacts of sentiment on the stock market is one of the main objectives of this thesis. In the process of achieving this objective, we explored different sources of sentiment information. We start with the processed and aggregated sentiment data at daily intervals to generate the sentiment information and examine its relationship with the stock market. We extend this work by developing a high-performing NLP algorithm that has been pre-trained on a large corpus of Wikipedia. Thereafter, the pre-trained model would be

trained on financial news data and the resulting trained NLP model would be tasked with extracting sentiment polarities from financial news and the results incorporated into our proposed event-driven stock market predictive models. Findings from this work would help to clarify the Granger causality between sentiment and the stock market. In addition to this knowledge, we would also use our models to predict the directions of the stock market.

Finally, the time sensitivity of the sentiment information with respect to its relevance on the stock market would be examined statistically with the aim of assessing the possibility of exploiting market imperfection due to information asymmetry.

1.2.2 Structure of the thesis and contributions

The work starts with a critically robust examination of the Granger causality between sentiment and the stock market. Of course, the first question that comes to mind would be: what is sentiment? Sentiment has very broad contextual implications. But for simplicity, we consider sentiments as emotional reactions or expressions to events communicated through news, blogs and/or other means of communications.

In our case we begin by assessing the Granger causality between the sentiments extracted from the public mood and the stock market. That is, our sentiment data is sourced from public blogs that are not focused on the stock market. This is addressed in Chapter 2 with a view to assessing if the sentiment data is influential in predicting the directions of the stock market.

We extend the work to also assess the influence of focused sentiment information on the stock market. In this case, we explore the sentiments that are related to the stock market index of interest and its constituents as detailed in Chapters 3 and 4. However, there is limited information in literature regarding the processing of the sentiment datasets explored in the chapters mentioned above. Another limitation was that the approaches used to construct these sentiment datasets, which are only described at the high level, were developed

before 2018 when a new powerful BERT-based NLP model was proposed in literature (Devlin et al. [69], Araci [21], and Olaniyan [111]). In light of these challenges, we developed a high-performing BERT-based NLP model with a high level of accuracy. The model is used to extract the sentiment polarities from the financial news dataset covered in Chapters 5 and 7.

In the previous chapters we utilised daily time series datasets - the sentiment and stock market variables are sampled at constant time interval of a day. We use intraday sentiment and stock market datasets in Chapters 6 and 8. One of the rationales for this choice is to examine the time sensitivity of the sentiment and the stock market variables. Also, we attempt to address the concerns with the use of daily time series such as poor statistical properties, among others (Easley et al. [24]).

For the visual illustration of the structure of the thesis, we have Fig. 1.2.1. Also, the subsections that follow detail the work conducted in each chapter.

Chapter	Sentiment	Stock	Comment
Chapter 1			Introduction.
Chapter 2	Public sentiments obtained from LiveJournal blogs	S&P 500 Index	Analysis of the Granger causality between sentiments and the stock market.
Chapter 3	Sentiments related to the S&P 500 index and sourced from Tweets.	S&P 500 Index	Analysis of the Granger causality between stock-related sentiments and the stock market.
Chapter 4	Sentiments related to the constituents of the S&P 500 and sourced from Quandl.	S&P 500 Index and the closing prices of its constituents.	This chapter extends the scope of the sentiment information explored in the previous chapters in that the sentiments gathered covered most of the constituents of the S&P 500.
Chapter 5	Financial news related to the constituents of the S&P 500 index.		High performing NLP model developed for extracting sentiment polarities from the financial news.
Chapter 6		E-mini S&P 500 - Intraday data	Intraday stock market data is introduced and explored.
Chapter 7	Sentiments from Chapter 5	Stock data from Chapter 6	Causality relationship between sentiments and the stock market
Chapter 8		Stock data from Chapter 6	Predicting model for trading Contracts for Difference (CFDs) is developed.

Figure 1.2.1: Structure of the thesis. Based on how the chapters relate to each other.

Chapter 2 - Social web-based anxiety index's predictive information on S&P 500 revisited

To investigate if sentiment has an impact on the stock market, this chapter begins by sourcing sentiment information from the public mood. In particular, the work of Gilbert and Karahalios [38] is revisited in this chapter. The aim is to examine the relationship between the stock market returns and the public mood obtained by extracting sentiment polarities from social blogs. A non-linear Granger causality approach is employed, in addition to the linear approach used by [38], to validate the Granger causality between them. But one still needs to exercise caution in reaching any conclusions. Clearly, there are some limitations in the research scope covered in Chapter 2. Can the findings from Chapter 2 be considered as a golden validation of the relationship between sentiment and the stock market? What are the asymmetric impacts of the positive and negative sentiments, if any at all? These important questions are answered in Chapter 3.

Chapter 3 - Sentiment and Stock Market Volatility Predictive Modelling - a Hybrid Approach

First, using the findings from Chapter 2 might not be sufficient enough to establish categorically the relationship between sentiment and the stock market as the sentiment data is sourced from the public mood. For example, one would hardly agree that some sentiments from completely random and unrelated public expressions would have an impact on the stock market. It is like saying the loud noise from an old slow-moving train might have an impact on the stock market. As a result of this concern, it is therefore important to use focused sentiment data that are directly related to the stock market to assess this relationship. That is, it would be wrong to generalise the relationship between sentiment and the stock market based on the findings from Chapter 2 simply because of the sentiment data source explored. Findings from Chapter 2 could only attempt to assess if the sentiment from the public mood has any statistical relevance on the stock market subject to the selection of appropriate models.

Chapter 3 would therefore extend the work by focusing on the sentiments related to the stock market as opposed to the public mood and assess the asymmetric impacts of the positive and negative sentiments on the stock market prices and volatility. The proposed non-linear Granger causality approach explored in Chapter 2 would be applied as well in Chapter 3.

Chapter 4 - Predicting S&P 500 based on its constituents and their social media-derived sentiment

To further assess the statistical significance of sentiment information on the stock market, the sentiments and the stock returns from the constituents of the S&P 500 are explored.

Chapter 4 is developed in light of the findings from the previous chapters by extracting the sentiments that are directly linked to the stock market indices of interest. This is achieved by generating the sentiments from news that contains any of the constituents of the S&P 500. This chapter is different from the previous chapters in that Chapter 2 focuses on the sentiments obtained from the public mood, and Chapter 3 uses the sentiments that are related to the stock market. The novelties of Chapter 4 begin with the peculiar nature of the data explored. At first, over 400 variables representing the closing stock prices of the S&P 500 constituents are pre-processed using clustering [18] and Principal Component Analysis (PCA) [94] in order to reduce the inherent dimensionality challenge. Then, we pre-process over 400 sentiment variables for the S&P 500 constituents. The combined results from these two exercises with lagging largely increase the dimensionality of the final data used. The use of a rolling window of 100 data points for the model development and 10 days for the forecasting further complicates the already complicated issue. In fact, to our knowledge, no other research work has used this dataset with lagging before. There are over 1200 variables to process after including lagging for every rolling window. Knowing that we could not include every variable in our predictive model development, we propose a variable selection model that allows us to identify the most significant

variables. This is another novelty of this chapter. The proposed variable selection model appears to have performed better than the popular Random Forest model. This piece of work takes an extra step to propose a strategically optimal trading algorithm based on the incorporation of machine learning and evolutionary optimisation algorithm [85] . This clearly shows another novelty of this chapter.

Chapter 5 - A two-step optimised BERT-based NLP algorithm for extracting sentiment from financial news

The research works from the previous chapters would attempt to evaluate the impacts of sentiment on the stock market. But the works seem to be limited in scope in that processed and aggregated sentiment datasets at a daily level are explored. The concern with the use of daily processed and aggregated sentiment data is that the stock market might have already incorporated any relevant new sentiment information into its prices, probably due to the time-sensitive nature of sentiment. To achieve market imperfection, some level of information asymmetry would be expected [113] and this requires us exploring news in real time as it becomes available.

As a starting point we aim to develop a NLP model. With this we can extract sentiments from raw financial news as opposed to using any refined sentiment data made available by third parties. A BERT-based NLP model that relies on the model pre-training for language transfer would be explored.

Chapter 6 - Event-based algorithmic intraday trading

We introduce an event-driven sampling of intraday financial data based on the time-varying return volatility. This ensures that only the informative features from the high volume of intraday stock market data are extracted for training our stock market predictive models. The aim is to develop a model for predicting the directions of the stock market. During this model development, we incorporate both the stock market data and technical analysis indicators.

In addition, we would introduce a novel approach for detecting

non-stationarity or white noise in variables and propose an optimal approach that fractionally differences variables in order to have them stationary as opposed to using a log differencing approach and compare the results from the models developed from these two stationarity approaches.

Chapter 7 - Event-based algorithmic intraday trading with applied natural language processing algorithms (NLP)

Chapter 7 would focus on the development of a systematic algorithm trading model that captures the events and trends in the stock market. As we are exploring high-frequency intraday stock market data, we would expect this to be highly voluminous. Clearly, we do not intend to inject all the data into the model - we sample more data during some events. Our methodology that detects the events in the market is detailed in the chapter.

There are two key questions to consider when defining events - are events driven by contagious sentiment or by the underlying stock index of interest itself? We apply our proposed optimised BERT-based NLP model developed in chapter 5 to extract sentiments from the raw financial news related to the constituents of the S&P 500 stock index. With this information we aim to examine the statistical relevance of sentiment on the stock market.

Chapter 8 - Contracts for Difference (CFDs) machine learning algorithm for optimising portfolios

In CFDs trading the chance of losing is very high. The Financial Conduct Authority (FCA) [59] has expressed some serious concerns that over 82% of people involved in the betting and CFDs trading lose money based on a sample of industry data. In light of this, expectations would be very high that any machine learning predictive models developed are statistically appropriate, perceptive, insightful, and have a high level of accuracy relatively before they become deployed for algorithmic trading. But there are frequent ups and downs and expected surprises in the stock market which could potentially weaken the

predictive power of the relevant models developed.

This chapter introduces a novel event-driven forward labelling approach for predicting the directions of the stock market based on the identified stock market events. In addition, a 2-step machine learning model based on the OXGBoost framework is employed.

Chapter 9 - Conclusion

This part summarises the work completed in each chapter of this thesis. It would include the contributions, constraints and limitations of the thesis and finally suggest future research work.

2

Social web-based anxiety index's predictive information on S&P 500 revisited

According to the investment theory [37], stock market is operating under the Efficient Market Hypothesis (EMH), in which stock prices are assumed to incorporate and reflect all known information. Sprenger et al. [130] strongly disagree with EMH by saying that the market is inefficient and therefore abnormal returns can be earned. In search of abnormal earnings, researchers now 'listen' to news and mine online aggregated social data all in the course for these attractive profits.

Schumaker and Chen [115] are among the early researchers to investigate whether emotions can predict the stock market. Machine learning algorithms

such as SVM, Naive Bayes, etc, are utilised to develop predictive models used to claim that financial news has a statistically significant impact on the stock market. Bollen et al. [66] present an interesting machine learning based approach to examine if emotions influence stock prices. Their results support the claim that emotions do influence the stock market.

The linear Granger causality analysis ([31],[109]) is employed by Gilbert and Karahalios [38] as a method to illustrate that web blog contained sentiment has predictive information on the stock market, but this method proved to have clear limitations as explained later in this chapter. A linear model and the Granger causality test are used also by Mao et al. [49] to examine the influence of social blogs on the stock market. The authors do raise some concerns about the possible non-linear nature in the relationship, but such concerns are not further explored. The non-linear Granger causality test, which relies on a Self-Organising Fuzzy Neural Network model, is unpopular in this area of work as it is thought not to be strong enough to capture volatile stock market movements, as revealed by Jahidul et al. [4]. Mittal and Goel [7] use machine learning algorithms to investigate if stock blogs, as a proxy for news, can predict this complex financial movement. Their findings make the same claim that stock blogs can be used to predict stock prices, and they use some level of accuracy of the predictive models to support their results.

The stock market is highly volatile. Therefore, capturing its movement and identifying relationships between stock prices and possible predictive variables require the use of appropriate approaches. These approaches should normally meet two requirements. The first requirement is to generate models for prediction, and the second requirement is to rigorously prove the models' predictive value.

As illustrated earlier in this section, there is a growing research work trying to establish that online expressed emotions have predictive information on the stock market. Most of these works fulfil the first requirement by devising and proposing various predictive models, but very few works attempt to fulfill also the second requirement by rigorously / statistically proving the predictive value

of these models. Gilbert and Karahalios [38] are among the very few that do consider both requirements, by proposing a statistical approach, which is based on the Granger causality analysis and Monte Carlo simulations. We recognise the large interest and potential generated by [38] in inspiring further research that demonstrates the link between the online expressed emotions and the stock market. Our work builds upon the approach presented in [38], and does so by critically analysing it, by clearly identifying its drawbacks and limitations, by tackling these limitations, and by extending the approach and the results presented in the chapter. As such, we establish our findings on data which has been obtained from the [38] 's authors website.

The remainder of this chapter is organized as follows. Section 3.1 briefly revisits the empirical analysis of Gilbert and Karahalios [38]. It presents the data, and the Anxiety Index's building process. In addition, we discuss the essential limitations of the approach of [38], and provide and discuss the results of our alternative Monte Carlo simulations. Section 3.2 presents our new statistical based approach which captures efficiently the stock market volatility, and the predictive information relationship direction between stock prices and emotion. Section 3.3 entails our findings and conclusion.

2.1 Discussion on the Web blog based Anxiety Index

Four stationary daily time series variables were explored in Gilbert and Karahalios [38]: the Anxiety Index (AI), the stock return, and two control variables which are the trading volume and the stock volatility. All the variables were generated from the stock market data S&P 500, except for the Anxiety Index AI.

[38] introduced the Anxiety Index using 20 million posts and blogs from LiveJournal, that had been gathered within three periods of 2008: January 25th to June 13th; August 1st to September 30th, and November 3rd to December 18th. Some of the examples of the raw public mood LiveJournal posts are provided below:

1. WELCOIN is a WELUPS MainNet token based on the TRC20 Token Standard and issued by NEEBANK - Digital Bank. In Welups - Blockchain and NFT platform, WELCOIN is a digital currency used for all trades, contracts, and services. Grab the opportunity to become a millionaire in the next few years.

2. Asian shares held onto their recent gains on Wednesday. The Shanghai composite is up 0.41% at 3,528.82. Overall, the Singapore MSCI up 0.21% at 357.75. Over in Hong Kong, the Hang Seng Index up 0.07% at 25,655. In Japan, the Nikkei 225 flat at 27,740, while the Topix index is up...

Two sets of linguistic classifiers trained with a LiveJournal mood corpus from 2004 were employed to build the Anxiety Index metric. First, a corpus of 624,905 mood′-annotated LiveJournal posts from Balog et al. [76] was used. 12,923 posts that users tagged as 'anxious', 'worried', 'nervous' or 'fearful' were extracted. Then two classifiers were trained to distinguish between 'anxious' and 'non anxious' posts. The first classifier $C1$, which was a boosted decision tree, as introduced by Yoav and Robert [44], used the most informative 100 word stems as features. The second classifier $C2$ consisting of a bagged Complement Naive Bayes model [29], used 46,438 words obtained from the 2004 corpus mentioned above. $C1_t$ and $C2_t$ were defined as the standard proportions of posts classified as 'anxious' by $C1$ and $C2$, respectively, during the closing trading day t. $C1_t$ and $C2_t$ were integrated in the series C defined by $C_t = max(C1_t, C2_t)$. The Anxiety Index was finally defined as the series $A_t = log(C_{t+1}) - log(C_t)$. 174 values were generated for this series from the available data.

The S&P 500 index was used as a proxy for the stock market and was employed to generate three variables participating in the development of predictive models, namely the stock market acceleration metric denoted as M, the return volatility denoted as Q, and the volume of stock trading denoted as V. The stock return at time t was defined as $R_t = log(SP_{t+1}) - log(SP_t)$, where SP is the closing stock price. The stock market acceleration metric was obtained from the stock return as $M_t = R_{t+1} - R_t$. The stock return volatility was expressed as

Table 2.1.1: Granger Causality results and Monte Carlo Simulation. $MCp_{Gausskern}$, MCp_{inv} and MCp_{boot} are the p-values of the simulations using a Gaussian kernel assumption, the inverse transform sampling, and bootstrap sampling respectively.

$F_{3,158}$	$p_{Granger}$	$MCp_{Gausskern}$	MCp_{inv}	MCp_{boot}
3.006	0.0322	0.045	0.045	0.045

$Q_t = R_{t+1} * R_{t+1} - R_t * R_t$, and finally V_t was expressed as the first difference of the lagged trading volume.

2.1.1 Findings and limitations

The two OLS models employed by Gilbert and Karahalios in [38] are:

$$M1 : M_t = \alpha + \Sigma_{i=1}^{3}\beta_i M_{t-i} + \Sigma_{i=1}^{3}\gamma_i V_{t-i} + \Sigma_{i=1}^{3}\delta_i Q_{t-i} + \varepsilon_t \tag{2.1}$$

$$M2 : M_t = \alpha + \Sigma_{i=1}^{3}\beta_i M_{t-i} + \Sigma_{i=1}^{3}\gamma_i V_{t-i} + \Sigma_{i=1}^{3}\delta_i Q_{t-i} + \Sigma_{i=1}^{3}\eta_i A_{t-i} + \varepsilon_t \tag{2.2}$$

The models $M1$ and $M2$ were used to measure the influence of the Anxiety Index on the stock prices. The difference in the models is that $M1$ does not include the Anxiety Index variable; it only uses the lagged market variables mentioned above in this section. $M2$ adds the lagged Anxiety Index to the $M1$'s variables. If $M2$ performs better than $M1$, one could conclude that the Anxiety Index has predictive information on the stock market. The first two columns of Table 2.1.1 show that $M2$, with the Anxiety Index included in the analysis, would outperform $M1$, judging from the Granger causality F statistics $F_{3,158} = 3.006$, and the corresponding p-value $p_{Granger} = 0.0322$.

The main disadvantage of the approach of Gilbert and Karahalios [38] was

that the Granger causality analysis's linear models M1 and M2 were actually not valid from a statistical point of view. These models suffered from major shortcomings, as, for instance residuals were non-normally distributed, and they presented a heterogeneity of the variance. As such, although the p-value $p_{Granger} < 0.05$ suggests that the Anxiety Index significantly adds some predictive information on the stock market; such a conclusion is not supported by a valid statistical reasoning.

Due to the mentioned pitfalls, [38] proposed also a Monte Carlo simulation with a Gaussian kernel distribution assumption for the Anxiety Index, in an attempt to retrieve the same conclusion as in the non-statistically supported Granger causality analysis. The authors generated 1 million sets of samples for the Anxiety Index. These new series were used in (2) by iterating 1 million times to generate the same number of F statistic values, and then to classify these values based on if any F statistic is at least 3.01. The total number of F statistic's values that were at least 3.01 was then divided by the number of iterations to obtain the Monte Carlo experimental p-value, $MCp_{Gausskern} = 0.045$, shown in Table 2.1.1.

Although $MCp_{Gausskern} < 0.05$ seemed to confirm the conclusion of the Granger causality analysis, the Monte Carlo simulation suffered at its turn of the issue of retrieving a significantly different experimental p-value with respect to $p_{Granger}$. This issue seemed to be the consequence of another issue, consisting of the fact that the empirical distribution of the F-statistic computed in the Monte Carlo experiments significantly deviated from the expected F-distribution, as confirmed by the Kolmogorov-Smirnov test, i.e. $D = 0.0337,\ p < 0.001$ [38].

This realization constitutes a nontrivial reason to question the Monte Carlo estimates, and a natural question which arises is: would the assumption of the Gaussian kernel distribution for the Anxiety Index have possibly introduced a bias in the simulation? To answer the question, we apply other non-parametric Monte Carlo simulation methods based on the inverse transform sampling method using the continuous version of the empirical distribution function corresponding to the original Anxiety Index's sample, and bootstrap sampling. We follow the same procedure as that used in [38]. Our Monte Carlo p-values are

presented in the columns four and five of Table 2.1.1, where MCp_{inv} and MCp_{boot} denote p-values issued from the use of the inverse transform sampling and the bootstrap sampling methods. Both simulations led to a similar value of 0.045. Moreover, in both cases the empirical distribution of the F-statistic computed in the Monte Carlo experiments is different from the expected F-distribution. These shortcomings confirm once again that proving the relationship between the Anxiety Index and stock prices is problematic if linear models are involved.

To this end, we propose a new statistical approach to solve the limitations in [38] and to also reveal the relationship direction between the variables of interest.

2.2 Anxiety Index's predictive information on the stock market, revisited

We follow the guidelines from Diks and Panchenko [16] (see [50] for detailed explanation and software) to examine the line of Granger causality between the variables involved in our analysis. The idea of the non-parametric statistical technique for detecting nonlinear causal relationships between the residuals of linear models was proposed by Baek and Brock [35]. It was later modified by Hiemstra and Jones [17] and this has become one of the most popular techniques for detecting nonlinear causal relationships in variables.

Consider two series X_t and Y_t as follows: let the Lx and Ly be the lag length of the lag series X_t^{Lx} and Y_t^{Ly} of X_t and Y_t respectively, and let us denote the k-length lead vector of Y_t by Y_t^k. In other words,

$$\begin{aligned} Y_t^k &\equiv (Y_t, Y_{t+1}, ..., Y_{t+k-1}), k = 1, 2, ..., t = 1, 2, .., \\ Y_t^{Ly} &\equiv (Y_{t-Ly}, Y_{t-Ly+1}, ..., Y_{t-1}), Ly = 1, 2, ..., t = Ly + 1, Ly + 2, ..., \\ X_t^{Lx} &\equiv (X_{t-Lx}, X_{t-Lx+1}, ..., Y_{t-1}), Ly = 1, 2, ..., t = Lx + 1, Lx + 2, ..., \end{aligned} \tag{2.3}$$

Given arbitrary values for $k, Lx, Ly \geq 1$ and $\varepsilon > 0$, then X_t does not strictly

nonlinearly Granger cause Y_t if:

$$\begin{aligned} Pr(\parallel Y_t^k - Y_s^k \parallel < \varepsilon \mid \parallel Y_t^{Ly} - Y_s^{Ly} \parallel < \varepsilon, \parallel X_t^{Lx} - X_s^{Lx} \parallel < \varepsilon) \\ = Pr(\parallel Y_t^k - Y_s^k \parallel < \varepsilon \mid \parallel Y_t^{Ly} - Y_s^{Ly} \parallel < \varepsilon) \end{aligned} \tag{2.4}$$

where $Pr(A \mid B)$ denotes the probability of A given B, $\parallel \cdot \parallel$ is the maximum norm, i.e. for a vector $V \equiv (v_1, v_2, \ldots, v_m)$, $\parallel V \parallel = max\{v_1, \ldots, v_m\}$, $s,t = max(Lx, Ly) + 1, \ldots, N - k + 1$, N is the length of the time series and ε is N-dependent and typically has values between 0.5 and 1.5 after normalising the time series to unit variance. The left hand side in (2.4) is the conditional probability which implies that two arbitrary k-length lead vectors of Y_t are within a distance ε, given that two associating Lx- length lag vector of X_t and two associating Ly-length lag vector of Y_t are within a distance of ε. The right hand side in (2.4) is the probability that two arbitrary k-length lead vectors of Y_t are within a distance of ε, given that the two corresponding Ly-length lag vector of Y are within the distance of ε.

Eq.(2.4) can be rewritten using conditional probabilities in terms of the ratios of joint probabilities as follows:

$$\frac{CI(k + Ly, Lx, \varepsilon)}{CI(Ly, Lx, \varepsilon)} = \frac{CI(k + Ly, \varepsilon)}{CI(Ly, \varepsilon)} \tag{2.5}$$

The joint probabilities are defined as:

$$\begin{aligned} CI(k + Ly, Lx, \varepsilon) &\equiv Pr(\parallel Y_t^{k+Ly} - Y_s^{k+Ly} \parallel < \varepsilon, \parallel X_t^{Lx} - X_s^{Lx} \parallel < \varepsilon), \\ CI(Ly, Lx, \varepsilon) &\equiv Pr(\parallel Y_t^{Ly} - Y_s^{Ly} \parallel < \varepsilon, \parallel X_t^{Lx} - X_s^{Lx} \parallel < \varepsilon), \\ CI(k + Ly, \varepsilon) &\equiv Pr(\parallel Y_t^{k+Ly} - Y_s^{k+Ly} \parallel < \varepsilon), \\ CI(Ly, \varepsilon) &\equiv Pr(\parallel Y_t^{Ly} - Y_s^{Ly} \parallel < \varepsilon) \end{aligned} \tag{2.6}$$

The Correlation-Integral estimators of the joint probabilities expressed in Eq.(2.6) measure the distance of realizations of a random variable at two different

times. They are proportions defined as the number of observations within the distance ε to the total number of observations. Let us denote the time series of realizations of X and Y as x_t and y_t for $t = 1, 2, \ldots, N$ and let y_t^k, y_t^{Ly} and x_t^{Lx} denote the k-length lead, and Lx-length lag vectors of x_t and the Ly-length lag vectors of y_t as defined in (2.3). In addition, let $I(Z_1, Z_2, \varepsilon)$ denote a kernel that equals 1 when two conformable vectors Z_1 and Z_2 are within the maximum-norm distance ε of each other and 0 otherwise. The Correlation-Integral estimators of the joint probabilities in equation (2.6) can be expressed as:

$$\begin{aligned} CI(k+Ly, Lx, \varepsilon, n) &\equiv \frac{2}{n(n-1)} \sum_{t<s} \sum I(y_t^{k+Ly}, y_s^{k+Ly}, \varepsilon) \cdot I(x_t^{Lx}, x_s^{Lx}, \varepsilon), \\ CI(Ly, Lx, \varepsilon, n) &\equiv \frac{2}{n(n-1)} \sum_{t<s} \sum I(y_t^{Ly}, y_s^{Ly}, \varepsilon) \cdot I(x_t^{Lx}, x_s^{Lx}, \varepsilon), \\ CI(k+Ly, \varepsilon, n) &\equiv \frac{2}{n(n-1)} \sum_{t<s} \sum I(y_t^{k+Ly}, y_s^{k+Ly}, \varepsilon), \\ CI(Ly, \varepsilon, n) &\equiv \frac{2}{n(n-1)} \sum_{t<s} \sum I(y_t^{Ly}, y_s^{Ly}, \varepsilon), \end{aligned} \tag{2.7}$$

where $t, s = max(Lx, Ly) + 1, \ldots, N - k + 1, n = N + 1 - k - max(Lx, Ly)$.

Given that two series, X and Y, are strictly stationary and meet the required mixing conditions mentioned in Denker and Keller [90], under the null hypothesis that X does not strictly Granger cause Y, the test statistics T is asymptotically normally distributed and it follows that:

$$T = \sqrt{n}\Big(\frac{CI(k+Ly, Lx, \varepsilon, n)}{CI(Ly, Lx, \varepsilon, n)} - \frac{CI(k+Ly, \varepsilon, n)}{CI(Ly, \varepsilon, n)} \Big) \sim N\big(0, \sigma^2(k, Ly, Lx, \varepsilon)\big) \tag{2.8}$$

where $n = N + 1 - k - max(Lx, Ly)$ and $\sigma^2(\cdot)$, the asymptotic variance of the modified Baek and Brock test statistics, and an estimator for it are defined in the Appendix in Hiemstra and Jones [17].

To test our variables for a possibly non-linear relation, we start by introducing

the general framework of our models. Consider a regression model with a constant conditional variance, $VAR(Y_t \mid X_{1,t}, \ldots, X_{m,t}) = \sigma_\varepsilon^2$. Then regressing Y_t on $X_{1,t}, \ldots, X_{m,t}$ can be generally denoted as:

$$Y_t = f(X_{1,t}, \ldots, X_{m,t}) + \varepsilon_t, \tag{2.9}$$

where ε_t is independent of $X_{1,t}, \ldots, X_{m,t}$ with expectation zero and constant conditional variance σ_ε^2. $f(\cdot)$ is the conditional expectation of $Y_t \mid X_{1,t}, \ldots, X_{m,t}$. Eq.(2.9) can be extended to include conditional heteroscedasticity as follows:

$$Y_t = f(X_{1,t}, \ldots, X_{m,t}) + \sigma(X_{1,t}, \ldots, X_{m,t})\varepsilon_t \tag{2.10}$$

where $\sigma^2(X_{1,t}, \ldots, X_{m,t})$ is the conditional variance of $Y_t \mid X_{1,t}, \ldots, X_{m,t}$ and ε_t has the mean 0 and the conditional variance 1. Since $\sigma(X_{1,t}, \ldots, X_{m,t})$ is a standard deviation, it is captured using a non-linear non-negative function in order to maintain its non-negative structure. This leads us to GARCH models [108]. Comparing Eq.(2.9) and Eq.(2.10), the first part of the right hand side of Eq.(2.9) is the same with that of Eq.(2.10). This is a linear model. The second part of the right hand side of Eq.(2.9) are residuals of the linear process. They represent the second part of the right hand side of Eq.(2.10). Eq.(2.9) can finally be presented in the VAR framework as:

$$\begin{aligned} M_t = c + \Sigma_{i=1}^{3} h_i M_{t-i} + \Sigma_{i=1}^{3} \gamma_i V_{t-i} + \Sigma_{i=1}^{3} \delta_i Q_{t-i} + \\ \Sigma_{i=1}^{3} \eta_i A_{t-i} + a_t \end{aligned} \tag{2.11}$$

$$\begin{aligned} A_t = c + \Sigma_{i=1}^{3} h_i M_{t-i} + \Sigma_{i=1}^{3} \gamma_i V_{t-i} + \Sigma_{i=1}^{3} \delta_i Q_{t-i} + \\ \Sigma_{i=1}^{3} \eta_i A_{t-i} + a_t \end{aligned} \tag{2.12}$$

Following the second part of the right hand side of Eq.(2.10), the residuals a_t

Table 2.2.1: Assigning values to ε, as of Diks and Panchenko [16]

n	100	200	500	1000	2000	5000	10,000	20,000	60,000
ε	1.5	1.5	1.5	1.2	1	0.76	0.62	0.51	0.37

from Eq.(2.11) and Eq.(2.12) are presented in GARCH(1,1) as:

$$a_t = \sigma_t \varepsilon_t \tag{2.13}$$

where $\sigma_t = \sqrt{w + a_1 a_{t-1}^2 + \beta_1 \sigma_{t-1}^2}$, in which w, a_1 and β_1 are constants. We finally derive the GARCH(1,1)-filtered residuals, standardized residuals, as

$$\varepsilon_t = \frac{a_t}{\sigma_t} \tag{2.14}$$

We obtain the residuals from the VAR model in Eq. (2.11) and (2.12). The test statistic in Eq. (2.8) is then applied to these residuals to detect the causal relation between the Anxiety Index and stock prices. Diks and Panchenko [16] provide some important improvement to the Non-linear Granger Causality test. [16] demonstrates that the value to be arbitrarily assigned to the distance ε is highly conditional on the length n of the time series. The larger the value n, the smaller the assigned value for ε and, the better and more accurate the results.

Most of the related works choose $k = Lx = Ly = 1$. The length of the series we are analysing is less than 200, so choosing ε=1.5 conforms with Table 2.2.1. Given $\varepsilon = 1.5$, $k = Lx = Ly = 1$, the results from the test are presented in Table 2.2.2.

Our first result in this framework seems to support the idea that the Anxiety Index has predictive information on the stock market, as this is based on the p-value of 0.017 shown in the first row of Table 2.2.2. Some re-considerations are necessary though.

Hiemstra and Jones [17] state that the non-linear structure of series is related

to ARCH errors. Anderson [127] proves that the volatility of time series contains predictive information flow. But Diks and Panchenko [16] warn that the presence of conditional heteroscedasticity in series could produce spurious results. To avoid any possible bias in our results, the residuals are applied to Eq.(2.13) to filter out any conditional heteroscedasticity in the residuals of the VAR models. We also rely on the GARCH(1,1)-filtered residuals to re-establish our findings.

We are able to identify, using the GARCH(1,1) results, that a_t from Eq.(2.11) is a GARCH process with ε_t being a Gaussian white noise (having the p-values $\alpha = 0.003,\ \beta < 0.001$ and Shapiro-Wilk = 0.383) and that a_t from Eq.(2.12) does not contain significant heteroscedasticity except that ε_t is an i.i.d. white noise with a heavy-tailed distribution (having the p-values $\alpha = 0.136,\ \beta = 0.454$ and Shapiro-Wilk = 0.018). We obtain GARCH(1,1)-filtered residuals and the test statistic in Eq.(2.8) is re-applied to three sets of residuals: OLS residuals from Eq.(2.11) and Eq.(2.12); GARCH(1,1)-filtered residuals of stock returns and OLS residuals from Eq.(2.12); and GARCH(1,1)-filtered residuals from both stock returns and Anxiety Index. The results we present in rows 2 and 3 of Table 2.2.2 show p-values > 0.05 and thus confirm that our earlier result presented in row 1 of Table 2.2.2 is biased by the presence of heteroscedasticity in the residuals. We are thus able to show that the Anxiety Index does not possess any significant predictive information on the stock market.

In view of our results above, we therefore claim that the conclusion from Gilbert and Karahalios [38] according to which the Anxiety Index has predictive information on the stock market is not valid, which is supported also by the fact that the statistical conditions to validate their results are not met.

2.3 Conclusion

This chapter proposes a new approach to statistically examine the relationship between the stock market and emotions expressed online. It proves that the Anxiety Index introduced by Gilbert and Karahalios [38] does not possess

Table 2.2.2: Non-linear Granger non-causality test

$AI => SP$		$SP => AI$	
Lx=Ly=1	p	**Lx=Ly=1**	p
Before filtering	0.017	Before filtering	0.182
$GARCH(1,1)_{SP}$	0.349	$GARCH(1,1)_{SP}$	0.922
$GARCH(1,1)_{SP,AI}$	0.718	$GARCH(1,1)_{SP,AI}$	0.685

predictive information with respect to the S&P 500. It does so by addressing the statistical limitations present in, and by extending the approach of [38].

The main drawback of the approach in [38] to proving the existence of the predictive information of the Anxiety Index with respect to the stock market was that this approach used a Granger causality analysis based on producing and assessing predictive linear models, which were actually not valid from a statistical point of view. These models suffered of major shortcomings as for instance residuals were non-normally distributed, and they presented a heterogeneity of the variance. In an attempt to partially correct the above shortcomings, the Monte Carlo simulation performed by assuming a Gaussian kernel based density for the Anxiety Index, was also biased as the empirical distribution of the employed F statistic significantly deviated from the expected F-distribution [38].

We note that Monte Carlo simulations using the Gaussian kernel density approach have their own bandwidth selection problem, which may bias the simulations - see Zambom and Dias [12]. We therefore re-designed the Monte Carlo simulation presented in [38] by using bootstrap samples of the Anxiety Index first, and the inverse transform sampling based on the continuous version of the empirical distribution function corresponding to the original Anxiety Index sample. The results showed no improvement. This re-confirms the non-linear nature in the relationship between the stock market and emotion, and the erratic volatility in the variables. Linear models appear to be too 'basic' to capture these complexities.

We have therefore extended the approach of [38] by proposing a more capable framework based on the non-linear models introduced in [16]. Our first result, based on a p-value of 0.017 obtained in the non-linear Granger non-causality test, capturing the predictive information of the Anxiety Index with respect to S&P 500, is biased by the presence of heteroscedasticity. We filtered out the heteroscedasticity in the residuals using Eq. (2.13) and our GARCH(1,1)-filtered residuals were used with the test statistic in Eq. (2.7). Our results, based on p-values > 0.05, express the true non-causality relationship of Anxiety Index with respect to S&P 500.

Although our work has established that the Anxiety Index does not have predictive information with respect to the stock market, by proposing a new approach which is statistically sound and more conclusive, there are still some concerns on how the Anxiety Index was built, based on incomplete data, non-specific LiveJournal posts, corpus challenges, non-representative data sample, among others. Further refining the process of defining the Anxiety Index by addressing the above-mentioned concerns, may help to fine-tune our empirical results and provide us with a more reliable predictive model.

3

Sentiment and stock market volatility predictive modelling - a hybrid approach

Standard finance models are built under the main assumption that investors act rationally. These models make use of conventional data like the stock market data. The models assume that stock market returns are equal to fundamental returns, where the market returns reflect all known information. In view of the assumptions of market efficiency and investor rationality, the Efficiency Market Hypothesis (EMH) became popular. This hypothesis adds substance to the traditional finance models as these reflect the idea that all new information has already been factored into the stock market prices. As shown in Chapter 2 the validity of the Standard finance models has become questionable because of their inability to capture the stock market trends. This has led to a new branch in stock market modelling - behavioural finance centred around sentiments.

Schumaker and Chen [115], Bollen et al. [66], Baker and Wurgler [88], among other, examined the causal relationship between sentiment and the stock market, and they all agreed that sentiment is statistically significant in stock market modelling.

So far, the causality relationship between the stock market and sentiment has been investigated. But there is little evidence to support that sentiments resolve stock market uncertainty: as we will show here, evidence rather indicates that sentiments induce volatility. How can we predict the impacts of sentiment on the stock market volatility? How can we investigate the asymmetric effects of different sentiments on the stock market volatility? Knowing that the GARCH framework is popular in predicting the stock market volatility, how can we develop a much more efficient stock market predictive model by using the GARCH model as a benchmark? These are the main questions this chapter focuses on.

Black [41] observed a negative correlation between current stock return and future return volatility because bad news tends to increase volatility as the realised return is lower than expected, and good news tends to reduce volatility as the realised return is higher than expected. Lee et al. [133] employed a generalized autoregressive conditional heteroscedasticity-in-mean specification to examine the impact of investment sentiments on stock return and volatility. They emphasized that focusing alone on the impact of sentiments either on the mean or variance in asset returns alone could lead to misspecification problems. A GARCH framework was used to analyse their effects, and results showed that shifts in sentiments are negatively correlated with market volatility. That is, volatility increases (decreases) when investors become more bearish (bullish).

Yanlin et al. [123] employed a nonlinear model that investigated the impact of sentiment-based information flow on the stock return volatility. A GARCH framework was introduced and results from their work supported that sentiments have statistical influence in predicting the stock volatility. This also attracts our interest as evidence from growing research work suggests that sentiments do not necessarily resolve uncertainty; rather, they induce volatility [123].

Mittal and Goel [7] used machine learning algorithms to investigate if stock blogs, as a proxy for news, can predict this complex financial movement. Their findings made the same claim that stock blogs can be used to predict the stock prices, and they used some level of accuracy of the predictive models to support their results.

In view of this area of growing interest, this thesis attempts to examine the relationship among stock market returns, volatility, and stock-related sentiments. Secondly, we investigate the asymmetric impacts of good and bad stock-related sentiments on the stock market volatility. More so, we propose a much more efficient volatility predictive model that incorporates both an EGARCH framework and an artificial neural network framework.

The remainder of this thesis is organized as follows: Section 3.1 describes the non-parametric approach we use, and presents our results of the causality relationship between sentiment and the stock market return. It also presents our benchmark volatility predictive model and assesses the asymmetric effects of good and bad sentiments on the stock market volatility. Section 3.2 entails our new hybrid approach that incorporates both the GARCH framework and the artificial neural network framework. Section 3.3 reveals our findings and concludes the thesis.

3.1 Stock market and sentiment

We use stationary daily time series variables obtained from stock market data and also stock-related sentiments to measure the influence sentiment has on the stock market returns. The S&P 500 index values from the 6th of September 2012 to 12th of May 2014 are used as a proxy for the stock market data and are employed to generate two variables participating in the development of predictive models, namely the stock market acceleration metric denoted as M and the volume of stock trading denoted as V. The stock return at time t is defined as $R_t = log(SP_{t+1}) - log(SP_t)$, where SP is the closing stock price. The stock market acceleration metric is obtained from the stock return as $M_t = R_{t+1} - R_t$. V_t is

expressed as the first difference of the lagged trading volume. The sentiment series S is obtained directly from StockTwits, which contains sentiment-filled S&P 500 blogs on Twitter (see the Downside Hedge website for more detailed explanation about the sentiment building process [34]). The sentiment data obtained has already been processed with the positive values representing positive sentiments and the negative values denoting negative sentiments. The examples of the StockTwits below represent the original stock-tweets:

1. $SPY Go down in Louisiana and chase down like five $10 million worth of deposits Monday or Tuesday.

2. Gold Stocks Very Oversold, But Need Macro Catalyst $GDXJ $HUI Also $GLD $SPX talkmarkets.com/content/com.

3. AABB AABBG.X Gold Exchange coming in September. AABBG.X Exchange will have 20 trading pairs allowing for cryptocurrency loans,gift cards,tied to Gold. AABB OTC has 100M+ in assets 72M Cash on hand 30M in Gold bullion. (TECHY) OTC has 9.46B OS & SP hit 99$ in February compared to AABB 2.3B OS less than 1/4 as many shares. Mr William Snyder owns 275M Shares and JUST RETIRED 120M of those shares, goes to show you his obscene confidence in AABB Mr Snyder also has strong connections to Barrick Gold. Have a Great weekend and remember Voyager Pre Exchange launch SP was .17 cents fast forward 5 months later Post Exchange launch SP hit 30$.

These examples represent the stock-related tweets that have been processed by StockTwits and explored in our work.

We now define $A_t = S_t - S_{t-1}$. Moreover, we include sentiment dummy variables so that we could measure the asymmetric impacts of positive and negative sentiments on the stock market volatility. We do not have access to these different sentiments. We resolve to using proxies for positive sentiment dummy variable $D_t = 1$ where $A_t - A_{t-1} > 0$ and 0 otherwise. We are able to generate the positive sentiment and negative sentiment series by defining $P_t = A_t^2 * D_t$

and $N_t = A_t^2 * (1 - D_t)$, respectively. Our volatility series Q is generated using the exponential GARCH(1,1), denoted also by EGARCH(1,1), as follows:

$$Q_t = ln(\sigma_t^2) = \omega + \beta ln(\sigma_{t-1}^2) + \alpha \left[\frac{\mid \varepsilon_{t-1} \mid}{\sigma_{t-1}} - \sqrt{\frac{2}{\pi}} \right] + \gamma \left(\frac{\varepsilon_{t-1}}{\sigma_{t-1}} \right) + \theta_1 P_{t-1} + \theta_2 N_{t-1} \tag{3.1}$$

where β measures the impact of past volatility on future volatility, α measures the impact of positive stock market shock on the stock volatility, γ captures the impact of negative stock market shock on the stock volatility, and θ_1 and θ_2 measure the impact of positive and negative sentiments, respectively, on the stock volatility.

3.1.1 Conventional Granger causality between sentiment and stock returns

In the process of determining the causal relationship between sentiment and stock market return we present these two OLS models:

$$M1 : M_t = \alpha_1 + \Sigma_{i=1}^{3} \beta_{1i} M_{t-i} + \Sigma_{i=1}^{3} \gamma_{1i} V_{t-i} + \Sigma_{i=1}^{3} \delta_{1i} Q_{t-i} + \varepsilon_{1t} \tag{3.2}$$

$$M2 : M_t = \alpha_2 + \Sigma_{i=1}^{3} \beta_{2i} M_{t-i} + \Sigma_{i=1}^{3} \gamma_{2i} V_{t-i} + \Sigma_{i=1}^{3} \delta_{2i} Q_{t-i} + \Sigma_{i=1}^{3} \eta_{2i} A_{t-i} + \varepsilon_{2t} \tag{3.3}$$

The models $M1$ and $M2$ are used to measure the influence of the sentiment on stock prices. The difference in the models is that $M1$ does not include the sentiment variable; it only uses the lagged market variables mentioned above in this section. $M2$ adds the lagged sentiment to the $M1$'s variables. If $M2$ performs better than $M1$, one could conclude that the sentiment has predictive information

on the stock market. But such a conclusion is dependent on the conditions under which the estimated residuals are normally distributed and homoscedastic in variance.

Before Eq. (3.2) and (3.3) can be estimated, the volatility series Q must be established and the influence of sentiment on volatility assessed. Does sentiment have predictive information on volatility? What asymmetric impacts do positive and negative stock market shocks have on volatility? What asymmetric impacts do good and bad sentiments have on volatility? Solving Eq. (3.1) and (3.3) provides answers to the questions.

The traditional volatility model is built using a GARCH approach that uses the residuals from a linear model as input to generate the volatility series.

We start by introducing the general framework of our models. Consider a regression modelling with a constant conditional variance, $VAR(Y_t \mid X_{1,t}, \ldots, X_{m,t}) = \sigma_\varepsilon^2$. Then regressing Y_t on $X_{1,t}, \ldots, X_{m,t}$ can be generally denoted as:

$$Y_t = f(X_{1,t}, \ldots, X_{m,t}) + \varepsilon_t, \tag{3.4}$$

where ε_t is independent of $X_{1,t}, \ldots, X_{m,t}$ with expectation equal to 0 and constant conditional variance σ_ε^2. Here $f(\cdot)$ is the conditional expectation of $Y_t \mid X_{1,t}, \ldots, X_{m,t}$. Eq. (3.4) can be extended to include conditional heteroscedasticity as follows:

$$Y_t = f(X_{1,t}, \ldots, X_{m,t}) + \sigma(X_{1,t}, \ldots, X_{m,t})\varepsilon_t \tag{3.5}$$

where $\sigma^2(X_{1,t}, \ldots, X_{m,t})$ is the conditional variance of $Y_t \mid X_{1,t}, \ldots, X_{m,t}$ and ε_t has the mean 0 and the conditional variance 1. Since $\sigma(X_{1,t}, \ldots, X_{m,t})$ is a standard deviation, it is captured using a non-linear non-negative function in order to maintain its non-negative structure. This leads us to the traditional GARCH model defined as:

$$\sigma_t^2 = \omega_0 + \beta_1 \sigma_{t-1}^2 + \alpha_1 \varepsilon_{t-1}^2 \tag{3.6}$$

Table 3.1.1: Only parameters from Eq. (3.1) that are statistically significant are reported. $LjungBox_R$ and $LjungBox_{R^2}$ denote Ljung-Box tests on the standardised residuals and squared residuals respectively.

Variable	**Estimate**	**t value**	**p-value**
ω	-2.1849	-3.7214	< 0.001
β	0.7690	12.4479	< 0.001
λ	0.2926	4.2497	< 0.001
θ_1	-6.1031	-2.3453	0.0190
Test	$LjungBox_R$	$LjungBox_{R^2}$	$ARCHLM$
p-value	> 0.05	> 0.05	> 0.05

The problem with Eq. (3.6) is that the asymmetric effects of different market shocks could not be captured. As a result, a new model was introduced by [22]. This model is called the Exponential GARCH model defined in Eq. (3.1) to capture these asymmetric effects of different shocks on the stock market volatility. This proposed model has earned popularity as it makes it possible to measure the asymmetric effects of market shocks. We use this model as our benchmark in predicting the stock market volatility.

In order to obtain the volatility series Eq. (3.3) is estimated without the variable Q and the model residuals are applied to Eq. (3.1) to generate Q. Table 3.1.1 presents the results of the estimated volatility model.

It is revealed that past volatility has a positive relationship with regard to future volatility. In fact, it is observed that it influences future volatility the most. It has been shown that negative market shocks are positively related to market return volatility. They increase the level of market risk and therefore influence the stock volatility positively. The asymmetric impacts of different sentiments on stock volatility are also captured. As it would be expected, positive sentiment reduces volatility. Oddly, negative sentiment does not appear to be statistically important. Goodness of fit tests are also employed on the standardised residuals and squared

residuals of the estimated EGARCH model. The insignificant p-values from the Ljung-Box tests on both the standardised residuals and squared residuals, and the ARCH LM test, suggest that the EGARCH model would fit the data well.

The volatility series obtained in Eq. (3.2) and (3.3) are estimated, and the linear Granger causality test results are presented in Table 3.1.2. The first two columns in the table show that $M2$, with the sentiment included in the analysis, would outperform $M1$, judging from the Granger causality F statistics $F_{3,401} = 6.5385$, and the corresponding p-value $p_{Granger} = 0.0003$.

Table 3.1.2: Granger Causality results and Monte Carlo Simulation. $MCp_{Gausskern}$, and MCp_{boot} are the p-values of the simulations using a Gaussian kernel assumption, and bootstrap sampling respectively.

$F_{3,401}$	$p_{Granger}$	$MCp_{Gausskern}$	MCp_{boot}	Shapiro-Wilk
6.5385	0.0003	0.0005	0.0005	0.0047

However, there are some concerns in the estimated models: the estimated residuals possess serious autocorrelation, are non-normal and heteroscedastic in variance (having p-values Ljung-Box < 0.05 for lags from 3, and Shapiro-Wilk $= 0.0047$) and the heteroscedastic presence is revealed in the EGARCH process in Table 3.1.1 (with p-value of $\beta < 0.001$). These are major shortcomings of the linear Granger causality test results according to which sentiment would be a determining factor in predicting the stock market returns. In an attempt to see if we could still rely on the test results, Monte Carlo simulations with a Gaussian kernel distribution assumption for the sentiment series are employed. 1 million sets of samples are generated for the sentiment and are fed into (3.3) by iterating 1 million times. The same number of F statistic values are generated in the process and then classified based on if the F statistic is at least 6.5385. The total number of F statistic values that are at least 6.5385 is then divided by the number of iterations to obtain the Monte Carlo experimental p-value $MCp_{Gausskern} = 0.0005$ as shown in the third column of Table 3.1.2.

Although $MCp_{Gausskern} = 0.0005$ seems to confirm the conclusion of the Granger causality analysis, the Monte Carlo simulation suffered at its turn of the issue of retrieving a significantly different experimental p-value with respect to $p_{Granger}$. This issue seems to be the consequence of another issue, consisting of the fact that the empirical distribution of the F-statistic computed in the Monte Carlo experiments significantly deviated from the expected F-distribution, as confirmed by the Kolmogorov-Smirnov test (D = 0.0348, $p < 0.001$).

This realization constitutes a nontrivial reason to question the Monte Carlo estimates, and a natural question which arises is: would the assumption of the Gaussian kernel distribution for the sentiments have possibly introduced a bias in the simulation? To answer this question, we apply another non-parametric Monte Carlo simulation method based on the bootstrap sampling. We follow the same procedure as that used in the Gaussian Kernel Monte Carlo simulation. The result is presented in the fourth column of Table 3.1.2, where MCp_{boot} denotes the p-value issued from the use of the bootstrap sampling method. The simulation led to a similar p-value of 0.0005. Also, the empirical distribution of the F-statistic computed in the bootstrap sampling Monte Carlo experiment is different from the expected F-distribution (Kolmogorov-Smirnov test result having D = 0.0351, $p < 0.001$). These shortcomings confirm once again that proving the relationship between the sentiment and the stock market is problematic if linear models are involved.

Although there are strong reasons to accept the Granger causality results, on one hand, there are also issues regarding the assumptions clearly stated under the linear regression modelling, such as the residuals must be independent, normally distributed, and homoscedastic in variance. All these assumptions are violated in our estimated models despite the fact that our Monte Carlo simulations fairly validate the Granger causality test results. As such, in the next subsection we devise a non-parametric non-linear Granger causality test in the context of our problem, in an attempt to overcome the limitations illustrated in the present subsection.

3.1.2 Extending the approach to non-parametric non-linear Granger causality

The stock market exhibits frequent volatility, and this makes linear frameworks less capable of capturing and predicting its trends. For stock market predictive values to be considered reliable, two key necessary and sufficient requirements must be met. The first would be to generate an acceptable predictive model and the second would be to prove the model's predictive value rigorously and statistically. The inability of any model to satisfy these two conditions casts doubt on its predictive value. This has been the case with most research work attempting to examine the causality direction between the stock market and sentiment-filled online expressions. Gilbert and Karaholios are among the very few that attempted to statistically prove their models' predictive value in their highly cited work [38]. But their results appeared to be biased as a consequence of their non-normal estimated model residuals and heteroscedasticity. These results have finally been proved not to be valid by further investigation in subsequent work [109].

In this section we apply the non-parametric statistical technique for detecting nonlinear causal relationships between the residuals of linear models, technique which was originally proposed by Baek and Brock [35] and was later modified by Hiemstra and Jones [17] to become one of the most popular techniques for detecting nonlinear causal relationships among variables. The technique is explained in detail in Chapter 2.

To resolve the shortcomings of the linear Granger causality test, VAR models for stock returns and sentiment are exploited. For stock market return, we make use of (3.3) and for sentiment model, we have:

$$\begin{aligned} A_t = c_3 + \Sigma_{i=1}^{3} h_{3i} M_{t-i} + \Sigma_{i=1}^{3} \gamma_{3i} V_{t-i} + \\ \Sigma_{i=1}^{3} \delta_{3i} Q_{t-i} + \Sigma_{i=1}^{3} \eta_{3i} A_{t-i} + \varepsilon_{3t} \end{aligned} \tag{3.7}$$

Table 3.1.3: Non-linear Granger non-causality tests. A and M are the sentiment and stock market returns, respectively. $A => M$, for example, denotes the Granger causality test with direction from A to M, i.e. sentiment predicts stock returns.

Lx=Ly=1	$p - value$
$A => M$	0.66433
$M => A$	0.30186

Note that (3.3) and (3.7) are estimated and the residuals from the estimated models are applied to (3.6).

The results of the tests presented in Table 3.1.3 show that sentiment does not have any predictive power on the stock market return, as the corresponding p-value of 0.66433 does not show statistical significance. This is clearly contrary to the findings of the linear Granger causality tests which have been invalidated by the presence of residual non-normality and heteroscedasticity.

Having observed no causal relationship between sentiment and stock market returns, can one reach the same conclusion that sentiment has no predictive power over the stock market volatility? Is the EGARCH model used for volatility models efficient in reliably predicting stock market volatility? We investigate these problems in the next section.

3.2 A hybrid approach to predicting stock volatility

In this section we will demonstrate the predictive power of sentiment on stock market volatility by proposing a hybrid approach based on the GARCH framework and the artificial neural network framework in which we consider feed-forward and recurrent neural networks. However, in order to propose this hybrid approach, we start by simply attempting to assess the predictive influence of sentiment on the volatility using the EGARCH framework alone first and evaluate the relative improvements when we enhance our approach with

feed-forward and Elman neural networks.

Monfared and Enke [117] recently proposed a hybrid approach that incorporated GJR GARCH and feed-forward neural networks (NNs) in predicting volatility. Their model was applied to conventional variables such as the market returns, and the variance of ten NASDAQ indices. Their findings showed that incorporating NNs into the GARCH framework improves volatility predictive performance. But how accurate is the GARCH framework employed in predicting the stock market? Can some non conventional variables like sentiment improve the performance of predictive models? We answer these questions by presenting new models that combine both EGARCH and neural network (NN) models.

Advancement in information processing technology contributed to the birth of NNs. According to Malliaris and Salchenberger [96], NNs present the relationship between the inputs and outputs using the architecture of human brain to process large information and detect patterns by interconnecting and organizing them in different layers for information processing purposes. These layers are formed by a set of processing elements or neurons. The layers are structured in a hierarchy consisting of input layers, output layers, and hidden layers. The connected nodes possess some weights which define the influence of the individual input cell to the output cell. These weights are extracted from the training data employed in the process of learning the relationship between the inputs and the outputs. Each of the processing elements is assigned an activation level, specified by continuous or discrete values. For neurons in the input layers, their activation levels are determined from the response obtained in the input signals within the environment. For neurons in the hidden or output layers, their activation levels are defined as a function of the activation levels of the neurons connected to them and the corresponding weights. The functions are called transfer functions, which may be in the form of a linear discriminant function with a value 1 for a positive signal if the value of the function exceeds a threshold level and 0 otherwise. The function may also be continuously nondecreasing, as is the case with the sigmoid functions. A feed-forward NN, for example, has a

Table 3.2.1: Correlation. *Res* denotes the response variable.

	Res	Q_1	Q_2	Q_3	P_1	P_2	N_1	N_2
Res	1.000	0.342	0.166	0.079	0.006	-0.045	0.028	0.075
Q_1	0.342	1.000	0.663	0.454	-0.437	-0.207	0.120	-0.070
Q_2	0.166	0.663	1.00	0.664	0.125	-0.437	-0.187	0.119
Q_3	0.079	0.454	0.664	1.000	0.069	0.123	-0.099	-0.187
P_1	0.006	-0.437	0.125	0.069	1.000	-0.110	-0.158	0.281
P_2	-0.045	-0.207	-0.437	0.123	-0.110	1.000	0.397	-0.157
N_1	0.028	0.120	-0.187	-0.099	-0.158	0.397	1.000	-0.105
N_2	0.075	-0.070	0.119	-0.187	0.281	-0.157	-0.105	1.000

one-directional signal flow mapping the inputs into the outputs from the input layer to the output layer. The applications of the NN family are very popular in areas such as classifications, predictions, and pattern recognition, among others.

The NN family have different parameters in their design and these parameters may alter their outputs. Therefore, they are designed for different research goals. The backpropagation NN is one of the most popular regarding the areas of research work aforementioned. Collins et al. [36] applied it to underwriting problems. Malliaris and Salchenberger [91] also applied the backpropagation network in estimating option prices. To determine the values for the parameters in the algorithms, mean square error and gradient descent are employed. At each iteration, current parameters are updated by minimizing the mean square error differences between the actual response values and desired response values. A detailed explanation of the process is provided by Rumelhart and McClelland [27]. Feed-forward, multilayer, and recursive NN, such as Jordan recursive NN and Elman recursive NN, have become popular and preferred to the traditional NN techniques. The relationship between input variables and response variable is learned during the data training process with networks learning repeated from previous examples and data.

We build our stock volatility predictive approach based on feed-forward and recurrent NN combined with EGARCH models. In order to assess the impact of sentiment on the performance of the prediction we consider two sets of input datasets. The first set contains only the lagged volatility series fitted by the EGARCH model, Q_{t-1}, and Q_{t-3}. The second dataset includes, in addition to the first dataset, lagged positive and negative sentiments P_{t-1}, P_{t-1}, N_{t-1} and N_{t-2}. Q_{t-2} is excluded in both datasets because it is highly correlated with Q_{t-1} as presented in Table 3.2.1.

We employ the series on various classes of NN models: the feed-forward NNs, the Elman recursive NNs and the Jordan recursive NNs. Knowing that the output of NN models is sensitive to the values assigned to the parameters in the models (including the number of hidden layers, the number of their nodes, and the weights), with some computational efforts optimised NN models have been generated, and the Root Mean Square Errors (RMSE) have also been obtained as presented in Tables 3.2.2 and 3.2.3. The Elman NN and the feed-forward NN provide closely the same results and are clearly better than the Jordan NN. The RMSE from the Feed-forward and Elman models in Table 3.2.3 are lower than their corresponding RMSE in Table 3.2.2. This observation confirms the importance of including sentiments among the predictors of stock market volatility. The RMSE are clearly diminished when sentiment variables are included in the training. This observation is further investigated by employing the graphical representations of the trained NN models.

Figure 3.2.1 presents the regression plots of our fitted volatility for the two datasets. The performance is judged by the closeness of the fitted volatility plot in red to the optimal line in black. The plots in the first column of Figure 3.2.1 represent charts from the dataset without sentiment variables, and the plots in the second column denote charts from the dataset with sentiment variables included. The difference between the two datasets used is clearly presented by the feed-forward and the Elman NN models. The plots in the second column

Table 3.2.2: Results from the use of the dataset without sentiment series. *Size* refers to the number of hidden units, *Max* denotes the number of iterations, *Weight* denotes the weight decay and *RMSE* is the root mean square error which is the square root of MSE.

NN	*Size*	*Max*	*Weight*	*RMSE*
Jordan	16	340		0.00017
Elman	24	1440		0.00010
Feed-Forward	29	1400	0.001	0.00011

Table 3.2.3: Results from the use of dataset with sentiment series. *Size* refers to the number of hidden units, *Max* denotes the number of iterations, *Weight* denotes the weight decay and *RMSE* is the root mean square error which is the square root of MSE.

NN	*Size*	*Max*	*Weight*	*RMSE*
Jordan	20	1240		0.00017
Elman	30	1040		0.00004
Feed-Forward	30	1920	0.001	0.00005

appear much better than those in the first column and this suggests that by including sentiment variables one produces better predictions. That is, sentiment plays an important role in predicting stock market volatility.

Substantial evidence shows that sentiment has predictive information on the stock market. The previously produced EGARCH model has shown the significant importance of individual predictors. The relative importance of individual input variables in predicting stock market volatility is also investigated now in the context of our proposed hybrid approach combining EGARCH and NNs models. The information contained in Figure 3.2.2 follows the same direction of interpretations as that presented in Table 3.1.1. Past volatility

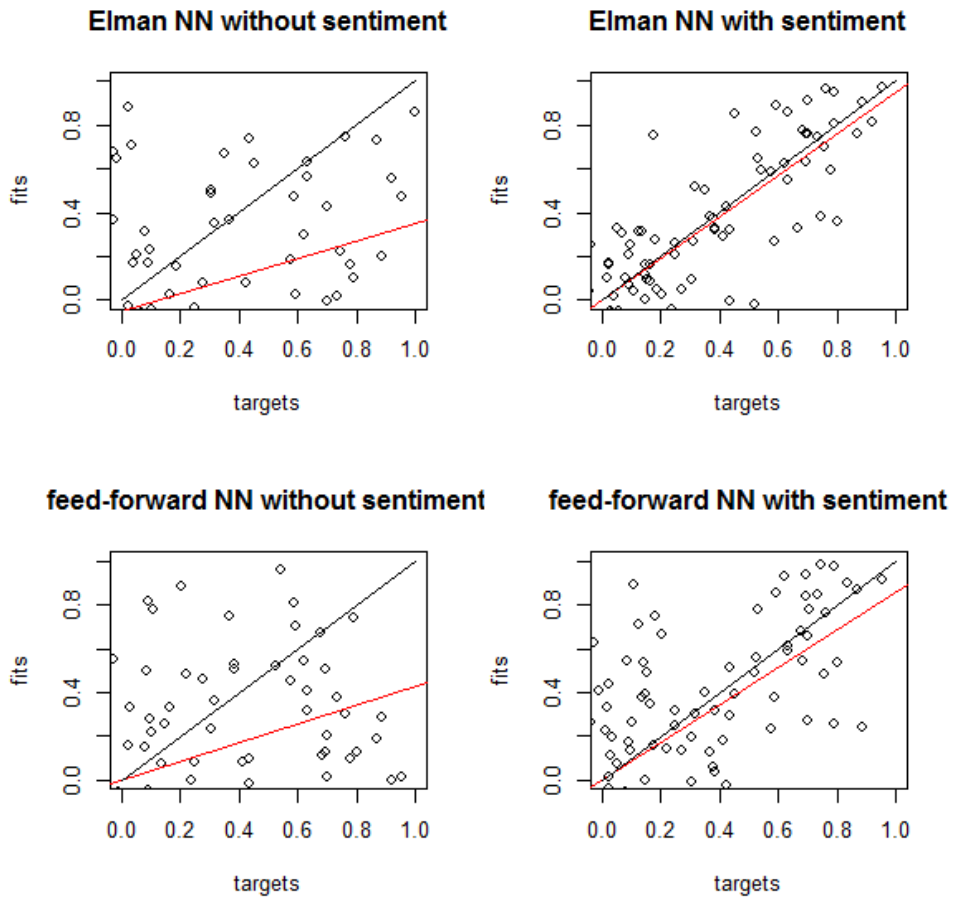

Figure 3.2.1: Regression model. It presents the information about the fitted volatility line in red and the optimal line in black. The figures in the first column represent plots of NN models with the dataset without sentiment variables. The figures in the second column represent volatility plots of the dataset with sentiment variables.

influences future volatility the most and it is positively related to future volatility. Positive sentiment has a negative relationship with future volatility. Of all the predictors, negative sentiments appear to have the least influence on the stock volatility.

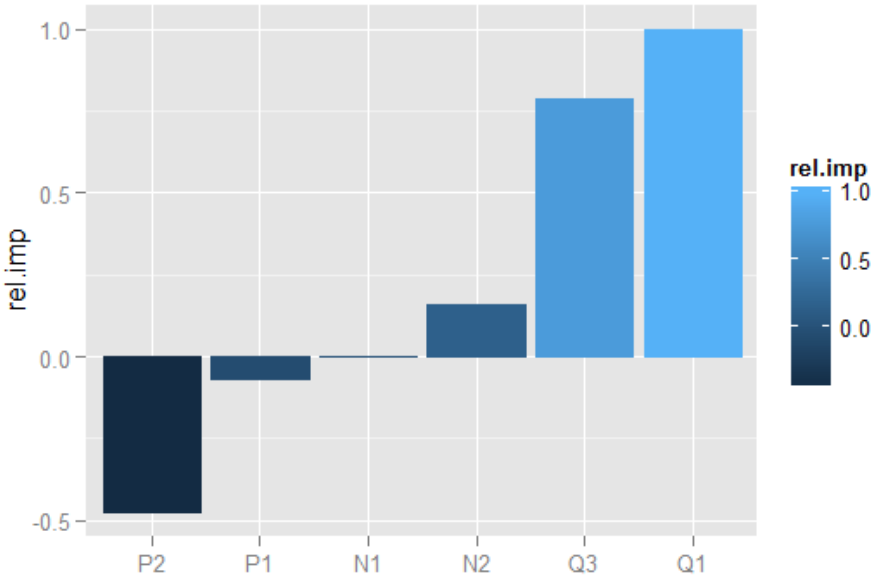

Figure 3.2.2: Relative importance. It measures the relative importance of the predictors in the model. Variables on the horizontal lines are the predictors. The variables with values below 0 have a negative relationship with the response variable and those with values above 0 have a positive relationship with the response variable. The response variable is the future volatility.

3.2.1 Sensitivity analysis

We have shown how individual variables impact on the response variables. Our findings present sentiment variables to be influential in predicting the stock market volatility. We have also shown the direction of the influence each variable has in predicting the stock market volatility. Recalling from the benchmark GARCH model used, past volatility has a positive impact on future volatility. Positive sentiment has a negative influence on the stock market volatility. From the relative importance information of the explanatory variables presented in Figure 3.2.2 it is observed that positive sentiment and past volatility have higher impacts on the volatility just the same way as presented from the results obtained in the GARCH model. The most relevant questions about the relationship between sentiment and the stock market variables have been answered from a

rigorous / statistical point of view. Equally important is the proposed hybrid approach that derives a larger efficiency from the combination of the GARCH framework and the neural network framework in developing more advanced volatility predictive models. We have shown that our proposed model is highly efficient in this regard. Yet, some important questions are still left unanswered.

For a basic linear regression model, it is easy to observe how each explanatory variable impacts on the dependent variable by keeping other explanatory variables constant. Secondly, it provides categorical information about the direction of relationships between individual explanatory variables and the dependent variable. Being categorical implies that a relationship shows if an explanatory variable has a positive or negative influence on the dependent variable, and that this direction of the relationship is constant. Clearly, linear models are simple, categorical, and straight-forward. This brings forward the question: is there any way to present the form of relationship of every individual explanatory variable with the response variable from the proposed hybrid approach? That is, given other explanatory variables constant, what amount of change will be impacted on the response variable for a unit change in an explanatory variable?

Neural networks are considered a 'black box' as they do not offer any insightful explanations about the impacts of individual input variables in the prediction process. Gevrey et al. [93] are among the early researchers that provided these long-awaited insights. This makes it possible to carry out sensitivity analysis on these individual explanatory variables. In order for us to answer the pressing question about the form of relationship of each explanatory variable on the dependent variable we use techniques of sensitivity analysis. We intend to examine how a unit change in each explanatory variable influences the dependent variable. We also aim to examine if the relationship between an explanatory variable and response changes with regard to the constant values of all other explanatory variables.

Figure 3.2.3 presents our sensitivity analysis results. Each of the 6 columns

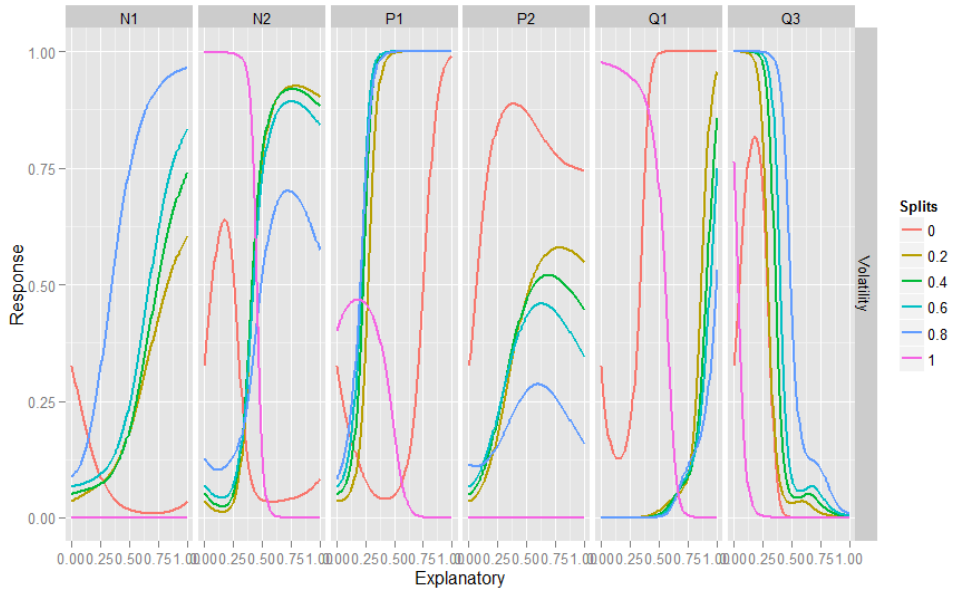

Figure 3.2.3: Sensitivity analysis plots. It depicts the forms of the relationship between each explanatory variable with regard to the dependent variable while keeping the other explanatory variables constant. N_1 and N_2 denote first and second lagged negative sentiment variables respectively. P_1 and P_2 denote first and second lagged positive sentiment variables respectively. Q_1 and Q_2 are first and third lagged volatility variables. The dataset is normalised to be between 0 and 1.

corresponds to each explanatory variable whose label is provided on top. At the far-right of Figure 3.2.3 we have *Splits*, which denotes the different constant values assigned to other explanatory variables, while one explanatory variable is under consideration with respect to the response variable. For each constant value there is a corresponding colour line as illustrated in the figure.

First, we determine the relative importance of the individual explanatory variables based on the slope of the curves. Starting with the first column with the label N_1, when other variables are kept constant at value 0, the topmost line denotes the relationship of the negative sentiment variable N_1 with respect to the response variable. It is observed that there is a negative relationship between these two variables, up until N_1 is 0.5. At this stage, the relationship is inelastic.

That is, less than a unit change is expected in the sentiment variable N_1 to cause a unit change in the response variable. From the point where N_1 is 0.5 and above, the form of the relationship changes to positive, and the relationship is elastic. This shows that the form of the relationship between the explanatory variable and the response variable may not necessarily be constant over time. In the second column corresponding to the negative sentiment variable N_2, we observe a positive relationship up to the point where the value of N_2 is around 0.2 when other explanatory variables are held constant at value 0. When it has values between 0.25 and 0.5, the direction of the relationship changes to negative and it is also inelastic, which means that less than a unit change in N_2 is expected to cause a unit change in the response variable. Above the value of 0.5, the relationship changes again. This also confirms that the form of the relationship using our hybrid approach is not constant and that may be the case with most neural network models. Comparing the two negative sentiment variables N_1 and N_2, all lines of N_2 are longer than those of N_1, which means that the relationships for N_2 are generally more inelastic (less elastic) than those for N_1. In terms of these variables' relative importance, N_2 is therefore more important than N_1. Interestingly, it is revealed that the form of the relationship between an explanatory variable with respect to the response variable differs for different constant values for all other explanatory variables. When the other explanatory variables are kept constant at the maximum value 1, column 1 shows a change in relationship and it reveals that N_1 does not have any influence on the response variable. At this stage, variables N_2 and Q_1 appear to be the most important predictors in relation to other predictors. When the constant values are set between 0.2 and 0.8, P_1 and Q_3 are the most important predictors with very strong inelastic relationships. Comparing all the explanatory variables relatively, P_1, Q_1 and Q_3 are the most influential predictors. Recalling from Figure 3.2.2 and the EGARCH results in Table 3.1.1 that show positive sentiment to be negatively related with volatility, the information presented in Figure 3.2.3 does not disprove this form of relationship. These results from Figure 3.2.2 and the estimated EGARCH model are retrievable under some constant values of other

explanatory variables.

This analysis underscores the importance of sentiment variables in developing stock market volatility predictive models. Considering stock-related online expressions, our sensitivity analysis supports that positive sentiment is significantly influential in predicting stock market volatility. We have shown also that individual explanatory variables do not necessarily have a constant form of relationship with respect to the response variable. Different values of other explanatory variables may cause an explanatory variable to change the form of the relationship it has with the response variable.

3.3 Discussion and conclusion

This thesis proposes a hybrid approach that incorporates a GARCH framework and feed-forward neural network model in developing a much more efficient volatility predictive model. It also details the relationship among sentiments, stock market returns and volatility by applying a non-parametric non-linear Granger causality framework to assess the causality direction between sentiment and stock market returns.

The linear Granger causality test shows that sentiment has a significant influence on the stock market returns. But the residuals obtained from the estimated linear model are non-normal and heteroscedastic in variance. These shortcomings invalidate the linear approach. But many related research work reach a conclusion based on the results from the linear approach without statistically validating their results as revealed by [109]. The violations of the linear assumptions about non-normal residuals and heteroscedasticity completely render the linear results invalid. Although Monte Carlo simulations based on the assumptions of Gaussian Kernel and Bootstrap sampling are also used to see if the experimental p-values obtained would provide some level of support. Well, the Monte Carlo results do support that sentiment does influence the stock market returns. As would be expected, the experimental F distribution from the Monte Carlo simulations deviates from the expected F-distribution.

This renders invalid the relationship between the Monte Carlo experimental results and our original linear results. In view of these shortcomings, the non-parametric nonlinear Granger test introduced by [109] was employed to examine the true relationship between the variables. The causality test reveals that there is no line of Granger causality between them. The linear model is biased as revealed by the violations of the key linear assumptions for validity. The strong relationship as shown in the linear Granger causality result disappears when the non-parametric nonlinear model is applied. It implies that the price of sentiments from stock-related blogs have already been factored into the stock market returns. The only logical reason one could reach is that since the information from stock-related blogs is public, then the news or blogs are not new so as to have any effect on stock market prices. The market is already considered efficient and therefore no advantage can be obtained from public information. Having concluded this, does it mean that sentiment does not influence the stock market at all? We move further to investigate the influence of sentiment on stock market volatility. In addition, we develop upon the existing stock market volatility predictive model by introducing a neural network framework into the traditional model.

Our volatility model built on the EGARCH framework shows that future volatility is influenced by factors such as past volatility and positive sentiments. Negative sentiment from stock-related news does not appear to have any influence on volatility. The RMSE obtained from the EGARCH model is 9.724997. This is used as a benchmark to compare the efficiency of our proposed hybrid model. The RMSE is reduced to 0.0005 by our proposed model. It clearly confirms the superiority of our model over the benchmark. The model also reveals the relative importance and directional influence of individual variables.

We are able to show the asymmetric impacts of positive and negative sentiments on the stock market volatility using both the conventional and our hybrid models. Positive sentiment influences volatility. This could be because the stock market is considered highly risky and therefore investors react to future volatility based on the information from past volatility. As such, substantial

positive sentiment is expected to cause changes in their investment portfolio.

We employ sensitivity analysis to examine the form of relationship each explanatory variable has with respect to the stock market volatility from our hybrid model. Our results show that the form of relation could be positive, negative, bi-modal or come in any kind of form. It all depends on the values of other explanatory variables employed at a point in time. Regardless, our sensitivity analysis shows that positive sentiment possesses predictive power on stock market volatility. It also shows that past volatility impacts future volatility.

In conclusion, we have shown that sentiment built-up process is a determining factor when measuring the effects of sentiments on stock market volatility. [109] uses sentiments that are not stock related. The findings from their work showed that past volatility and negative sentiments influence stock market volatility. Positive sentiment does not have any significant impact on volatility. But when sentiments are generated from stock-related blogs, past volatility still appears to have the strongest effect on future volatility. Positive sentiment has more effects on stock market volatility than negative sentiment. This implies that the source of sentiments used also has importance and therefore one must pay attention to the source of sentiments used in developing stock market predictive models.

4

Predicting S&P 500 based on its constituents and their social media derived sentiment

4.1 Motivation

We examined the Granger causality relationship between sentiment and the stock market in chapters 2 and 3 but with some variant based on the different sources of sentiment data. Interestingly, both chapters agree that sentiment does not significantly influence the stock market returns. But Chapter 3 reveals that it does impact the stock market volatility. Yet, attempts made so far have not yielded any objective conclusion as to this relationship. On this account we approach this sensitive subject by exploring sentiment directly sourced from the

constituents of the S&P 500 in examining this causal relationship between sentiment and the stock market.

4.2 Introduction

The influence of sentiments on the stock market has been extensively studied and so are the asymmetric impacts of positive and negative news on the market. But little has been done in devising efficient predictive models that can help to maximise investment portfolios while taking into consideration the statistical relevance of sentiment, and the proposed work addresses this concern. The main aim of this chapter is to predict reliably the directions of the S&P 500 closing prices, by proposing a predictive modelling approach based on integrating and analysing data on S&P 500 index, its constituents, and sentiments on these constituents. Indeed, this study is the first work to use constituent sentiments and its closing stock prices containing over 800 variables (combined closing stock prices of the S&P 500 constituents (Appendix A.0.2) and their respective sentiment data (Appendix A.0.3) without taking into account lagging - which further increases data dimensionality in a n-fold fashion) to predict the stock market.

First, we tackle the data high dimensionality challenge by devising and proposing a method of selecting variables by combining three steps based on variable clustering, PCA (Principal Component Analysis) [94], and finally on a modified version of the Best *GLM* variable selection method developed by McLeod and Xu [6].

Then we propose an efficient predictive modelling approach based on Jordan and Elman recurrent neural network algorithms. To avoid the pitfall of a time invariant relationship between the response and the explanatory variables in the highly volatile stock market data, our approach captures the dynamic of the explanatory variables set for every rolling window. This helps to incorporate the time-variant and dynamic relationship between the response and explanatory variables at every point of the rolling window using our variable selection

technique mentioned above. Finally, we propose an efficient hybrid trading model that incorporates a technical analysis, and machine learning and evolutionary optimisation algorithms [15].

We prove that our constituent and sentiment based approach is efficient in predicting S&P 500, and thus may be used to maximise investment portfolios regardless of whether the market is bullish or bearish [1]. This study extends our previous recent work on XLE index constituents' social media based sentiment informing the index trend and volatility prediction [43].

The remainder of this chapter is organized as follows. Section 4.3 presents our data pre-processing methodology, which is our proposed method for handling the data high-dimensionality challenge outlined above, for selecting the variables with predictive value. Section 4.4 elaborates on the results of the causality relationship between sentiment and the stock market returns using special techniques of Granger causality. Section 4.5 presents the predictive modelling approach that we propose based on machine learning techniques including Jordan and Elman recurrent Neural Network algorithms. Section 4.6 entails our proposed trading model that combines a technical analysis strategy and the estimated results from the machine learning framework to optimise investment portfolios with evolutionary optimisation techniques. Finally, Section 4.7 discusses our findings and concludes this thesis.

4.3 Stock data and sentiment information

In order to develop our approach to predicting the S&P 500 close prices, we rely on three main datasets which we integrate. The first dataset involves the collection of all the closing stock prices for the S&P 500 constituents obtained directly from Yahoo Finance website [58]. The second dataset is sentiment data for the constituents of the S&P 500 index, obtained from Quandl. It collects the contents of over 20 million news and blog sources in real time. These contents are

[1] Bullish and bearish are terms used to characterize trends in the stock markets: if prices tend to move up, it is a bull market; if prices move down, it is a bear market.

similar to the examples provided in Chapter 3. They retain the relevant articles and extrapolate the sentiments. The sentiment score is generated via its proprietary algorithm that uses deep learning, coupled with a bag-of-words and n-grams approach. According to Quandl, negative sentiments are rated between -1 and -5, while positive sentiments are rated between 1 and 5 ([55]). And the third dataset contains the S&P 500 historical close prices and trading volume, obtained also from [58].

All the data collected covered the period from 8th of February 2013 to 21st of January 2016. For S&P 500 and its constituents, the stock market return at time t is defined as $R_t = log(SP_{t+1}) - log(SP_t)$, where SP is the closing stock price. The stock market acceleration metric is obtained from the stock market return as $M_t = R_{t+1} - R_t$. Moreover, V_t is expressed as the first difference of the logged trading volume. Finally, the sentiment acceleration metric is defined as $A_t = S_t - S_{t-1}$, where S_t represents sentiment for each constituent of the S&P 500 at moment t.

By combining the three datasets, in all we have more than 800 initial variables to explore (not including lagged variables), which will lead to one of the challenges encountered in our framework in terms of high data dimensionality.

Figure 4.3.1 shows the data pre-processing process flow. It highlights all the processes undertaken to refine the data.

To handle the high data dimensionality challenge, we propose an approach to reducing the number of dimensions, adapted to our framework, based on 3 steps, consisting consecutively of performing variable clustering, PCA, and by applying a variable selection method that we introduce here based on Best *GLM* variable selection method developed by McLeod and Xu [6]. These steps are described in the following subsection.

4.3.1 Reducing data dimensionality

As mentioned, the prices and sentiments of the S&P 500 constituents are the variables of two of the datasets we dispose of initially. For analysis, it is important

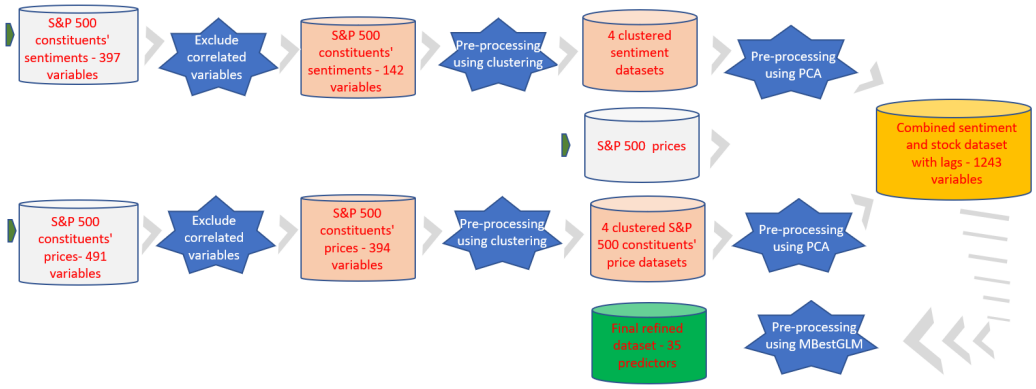

Figure 4.3.1: Data pre-processing process flow that details the processes followed to tackle the complexity of high dimensionality dataset. The three gray boxes represent the three main datasets. Detailed results of the K-means cluster are presented in Appendixes A.0.4 and A.0.5 for the constituents' stock prices and sentiments respectively.

to classify each constituent into groups. Of course, classifying the constituents based on their respective industries would have been the easiest way to group them since predefined information is readily available. Instead, we follow a more analytically rigorous approach in grouping the constituents based on pattern recognition and similarities in time series by using clustering.

In our case we use K-means clustering on each of the two sets of S&P 500 constituents for closing prices and sentiments respectively in order to group the variables in clusters. On the other hand, as we intend to use a rolling window of 100 days for testing and 10 days for forecasting, we note here that clustering is therefore applied on each rolling window, by forming 4 clusters. We note that by exploring different numbers of clusters between 3 and 10 on sample sets of 100 days rolling windows, 4 appeared to be the optimal number of clusters in all cases. Due to the generic property of within cluster similarity, it is expected that variables in a cluster are more or less similar.

When it comes to reducing the dimensionality of a numeric dataset, one of the most used methods is the well known Principal Component Analysis (PCA) [94].

With the advancement in data analytics, researchers and investment analysts have been able to explore big data in the process of developing predictive models. This helps to reduce the possibility of excluding relevant variables. But this benefit does not come at no cost. Imagine a situation whereby more than 500 variables are collected. Including all these variables in the final model might weaken the predictive power of the developed model. Then there comes the choice of selecting the variables that are statistically significant. And again, there is a likelihood that in the process of the selection, some important variables are ignored. This is the situation with high-dimensional data. In devising the means to solving this problem, various dimensionality reduction techniques have been considered by many researchers. Maaten et al. [87] explains dimensionality reduction as follows. Assume there is a dataset denoted as $n \times D$ matrix X with data vectors $x_i (i \in \{1, 2, 3, \ldots, n\})$ and dimensionality D. If it is further assumed that the dataset has some intrinsic dimensionality denoted as d where $d < D$, then the intrinsic dimensionality implies that the points in the dataset X are lying on or near a manifold with the dimensionality d which is within the D-dimensional space. Principal Component Analysis (PCA) is one of the most popular dimensionality reduction algorithms. It assumes that a number of observed variables can be reduced to a smaller number of artificial, called principal components, and still capture most of the variance of the observed variables. Hotelling [48] describes PCA as a linear technique for dimensionality reduction by embedding the data into a linear subspace of lower dimensionality.

Instead of applying PCA on all variables at once, we apply it to the groups of variables corresponding to each of the 4 clusters. When we combine the principal components from both sentiments and closing prices, we are still faced with a high number of dimensions of the combined dataset. By lagging the combined dataset up to 3 lags, its dimensionality increases to 1243 variables as shown in Fig. 4.3.1, which keeps the intended predictive modelling at a challenging level computationally and from a predictive modelling point of view. This led us to propose a variable selection method to handle this complexity in our approach.

As random forest is a popular technique used in variable selection, it was our

first choice to consider in order to further reduce data dimensionality in this 3rd step of our approach. We use a random forest model with default hyper-parameter values to select the best features. Interestingly, this solution performed very poorly on our dataset, judging by the poor goodness of fit with an Adjusted R-Squared being below 0.3 and 6 significant variables - 4 of these variables have 1% level of statistical significance and the remaining 2 variables have 0.5% level of statistical significance. We therefore proposed an alternative solution which is based on the modified best GLM method below developed by McLeod and Xu [6]. The latter selects the best subset of inputs for the *GLM* family. Given output Y on n predictors $X_1, \ldots, X_n$, it is assumed that Y can be predicted using just a subset $m < n$ predictors, $X_{i,1}, \ldots, X_{i,m}$. The aim is therefore to find the best subset of all the 2^n subsets based on some goodness-of-fit criterion. Consider a linear regression model with a number of t observations, $(x_{i,1}, \ldots, x_{i,n}, y_i)$ where $i = 1, 2, \ldots, t$. This may be expressed as

$$M_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_n x_{i,n} + \varepsilon_i \tag{4.1}$$

It is clear that when n is large, building 2^n regressions becomes computationally too expensive, and even untractable in our case with $n > 1200$ predictors, as mentioned above. As such, we modify McLeod and Xu's method of [6] as follows, and we call the resulting method MBestGLM. First, the lagged dataset is divided into subsets whereby each subset contains 35 predictors, and then the variable selection technique of [6] is applied on each subset with the intention of obtaining statistically significant predictors from each subset. The statistically significant predictors are then combined and the process of dividing the result into other subsets and applying the variable selection technique continues until the set of predictors can no longer be reduced. The regression results from these selected predictors produce a high adjusted R-Squared of over 0.65. The final dataset we obtain has an average number of predictors of 35. Indeed, from experiments we have seen that its number of predictors varies between 30 and 40 according to the instance of the rolling window on which the dataset is generated.

Overall, the dimensionality reduction process including the 3 steps of variable clustering, PCA, and the MBestGLM method we introduced above, are repeatedly applied on the rolling window as we work under the more general and thus more complex assumption of time-variant relationship between independent variables and return.

4.4 Sentiment's predictive information on S&P 500

As mentioned in the Introduction, it has been shown in a series of studies that sentiment variables help improve stock market prediction models ([88], [38], [66] and [110]). In light of this, it becomes imperative for us to investigate if the sentiment variables of S&P500 constituents included in our framework have some significant predictive power on this stock index.

In examining the relevance of sentiment variables, we use two methods, the first based on linear models, and the second one, more general, based on non-linear non-parametric models, respectively. These are Granger causality statistical tests and are used to see if sentiment has predictive information on S&P 500 in our framework.

4.4.1 Granger causality test: the linear model

Using the linear model framework represented by the Granger causality statistical test [20], we examine the causal relationship between sentiment and stock market returns. According to [20] we write the general linear VAR models as:

$$Model1 : M_t = a_1 + \Sigma_{i=1}^{3} \omega_{1i} M_{t-i} + \Sigma_{i=1}^{3} \beta_{1i} Stock_{t-i} + \varepsilon_{1t} \tag{4.2}$$

$$Model2 : M_t = a_2 + \Sigma_{i=1}^{3} \omega_{2i} M_{t-i} + \Sigma_{i=1}^{3} \beta_{2i} Stock_{t-i} + \Sigma_{i=1}^{3} \gamma_{2i} Sent_{t-i} + \varepsilon_{2t} \tag{4.3}$$

where M_t is the response variable which is the S&P 500 stock market return at

Table 4.4.1: Linear Granger causality results. $AdjR^2_{M1}$, and $AdjR^2_{M2}$ are the adjusted R-squared for $M1$ and $M2$ respectively. *pGranger* is the p-value for the Granger causality test between sentiment and the stock market.

$AdjR^2_{M1}$	$AdjR^2_{M2}$	$F_{16,165}$	*pGranger*
0.4319	0.6697	9.1439	< 0.0001

time t, M_{t-i} is the lagged S&P 500 market return with lag period of i, and *Stock* and *Sent* are variables generated by our 3-step dimensionality reduction process from the stock components and sentiment variables respectively. These VAR models *Model*1 and *Model*2 are used to examine if sentiment influences the stock market in our setting. As observed in the two equations, *Model*1 uses the lagged stock market return and the lagged stock market return principal components generated from the close prices of the S&P 500 constituents. In *Model*2 the lagged principal components, generated from sentiment variables related to the S&P 500 constituents, are added to the variables used in *Model*1. That is, *Model*1 does not contain sentiment variables while *Model*2 does. Sentiment variables would be considered to be influential if *Model*2 outperforms *Model*1 in prediction performance based on the adjusted R-squared metric. This is checked by using the standard Granger causality statistical test [20]. We consider the hypothesis Ho that *Model*2 does not outperform *Model*1, and we reject it by obtaining a significant p-value.

Our results, presented in Table 4.4.1, show that *Model*2, with the sentiment included in the analysis, outperforms *Model*1, based on the Granger causality F statistics $F_{16,165} = 9.1438$, and the corresponding p-value $p_{Granger}$ 0.05, and Shapiro-Wilk > 0.05), so the Granger causality test was applied correctly, and its conclusion is valid. Thus, sentiment has predictive information on S&P 500. In the next subsection we verify this conclusion with a more general non-parametric non-linear Granger causality test.

4.4.2 Granger causality test: the nonlinear model

The causality test from the linear model has already shown that sentiment variables have predictive power on the stock market. And the robust tests confirm that the results are not biased by the presence of autocorrelation or heteroscedasticity. Still, we examine the influence of sentimental information on the stock market using a non-linear non-parametric test which was originally proposed by Baek and Brock [35] and was later modified by Hiemstra and Jones [17].

Interestingly, the significant p-values from the nonlinear non-parametric technique (see [16] and [56] for detailed explanation and software used respectively) displayed in Table 4.4.2 prove that sentiment has predictive power on the stock market.

Table 4.4.2: Nonlinear non-parametric Granger tests. A and M are the sentiment and stock market returns respectively within the period 12/07/2013 and 16/05/2014. $A => M$, for example, denotes the Granger causality test with direction from A to M, i.e. sentiment predicts stock market returns. Similarly, $M => A$ is a Granger causality test if the stock market predicts sentiment.

$Lx = Ly = 1$	$p - value$
$A => M$	0.0077
$M => A$	0.0103

As a conclusion of this section, we can confidently state that the inclusion of sentiment variables does improve significantly stock market predictive models in terms of prediction performance, in our framework. Another interesting finding based on the significant p-value of $M => A$ in the nonlinear non-parametric Granger causality test, reveals that the stock market Granger-causes sentiment in this framework of S&P500 with its constituents and their sentiment.

4.5 Jordan and Elman neural network based approach to predicting S&P500 with sentiment

Linear and nonlinear models have been employed to assess the influence of sentiments on the stock market, and results have shown the statistical significance of sentiments' influence on the stock market in our setting. A linear model has also been developed in the previous section (see *Model2*) to investigate if the future S&P 500 close prices can be predicted with sentiment.

Neural Networks are predictive models capturing the relationship between inputs and outputs using a computational architecture inspired of the human brain, to process large information and detect patterns by interconnecting and organizing them in different layers for information processing purposes (Malliaris and Salchenberger [96]). These layers, hierarchically structured to consist of an input layer, an output layer, and hidden layers, are formed by a set of processing neurons. These layers are linked together by connected nodes and in-between the output layer and the input layer are the hidden layers which act as intermediaries. The connections between nodes possess some weights which define the strength of these connections. These weights are determined from the training data employed in the process of learning the relationship between the inputs and the outputs.

Each of the processing elements is assigned an activation level, specified by continuous or discrete values. For neurons in the input layers, their activation levels are determined from the response obtained in the input signals within the environment. For neurons in the hidden or output layers, their activation levels are defined as a function of the activation levels of the neurons connected to them and the corresponding weights. The functions are called transfer functions which may be in the form of a linear discriminant function with a value 1 for a positive signal if the value of the function exceeds a threshold level and 0 otherwise.

This section evaluates the relative improvements to the linear model when we enhance our approach by using Recurrent Neural Networks algorithms, more specifically for Jordan and Elman networks. The backpropagation algorithm is

one of the most popular techniques for training Neural Networks. It has been used in research works such as Collins et al. [36] which applied it to underwriting problems. Malliaris and Salchenberger [91] also applied backpropagation in estimating option prices. To determine the values for the parameters in the algorithms, the gradient descent technique is mostly employed Rumelhart and McClelland [27]. Multilayer, feed-forward, and recurrent Neural Networks such as Jordan and Elman Neural Networks which are used in this study, have become very popular.

As the datasets explored in our framework are highly dimensional, we rely on our variable selection methodology that we proposed in Subsection 4.3.1, to assist in selecting a reduced subset of variables based on S&P 500 index, its constituents and their sentiment, to implement a predictive modelling approach with Elman and Jordan Neural Network algorithms. That is, the same variable selection process used to obtain results from the estimated linear model in Section 4.4, is also used with our Neural Network models. It is important to note that in our approach we use a rolling window of 100 days for model development and fitting, and a rolling prediction period of 10 days. This choice was made based on several experiments we ran with our approach.

Table 4.5.1: Linear model (Linear), Jordan Neural Network (Jordan NN) and Elman Neural Network (Elman NN): The information represents the mean square errors in the period 11/08/2014 and 08/12/2014 based on the combined results from consecutive rolling windows.

	Linear	Jordan NN	Elman NN
Mean Square Error	0.0001092	0.0000980	0.0000993

Knowing that the output of Neural Network models is sensitive to the values assigned to the parameters in the models (including the number of hidden layers, the number of their nodes, and the weights), with some computational efforts, fairly optimised Neural Network models have been generated. Since at each

rolling window we may have a different selection of the set of predictors, the values assigned to Neural Network parameters would therefore be expected to be different for each fairly optimal result.

As observed in table 4.5.1, the Jordan and Elman Neural Network algorithms capture the stock market close price better than the linear model.

4.6 Evolutionary optimised trading model

In the previous sections we have demonstrated that sentiments influence stock market prices based on the results from the linear and Neural Network frameworks. But with all the information we have so far, are we able to maximise our investment portfolio by leveraging on the insightful information from our estimated models? We note that the information available still looks raw and therefore needs refining before we could make good use of it. In the process of refining the information, we resolve to introduce some stock market technical analysis and an evolutionary optimisation algorithm to our developed model. In doing so we propose the following strategies:

1. Active investment in *put* option with the expectation that price will fall in the future. The investor therefore profits from the fall in price. This helps to exploit bearish market.

2. Active investment in *call* option with the expectation that price will rise in the future. The investor therefore profits from the rise in price.

3. Hold position which implies that no investment should be made.

4. Passive investment refers to investment in stock market for a period of time without any optimal investment strategy.

Points 1 - 3 will be used to maximise investment portfolio under active investment and point 4 will be used to compare active and passive investment strategies.

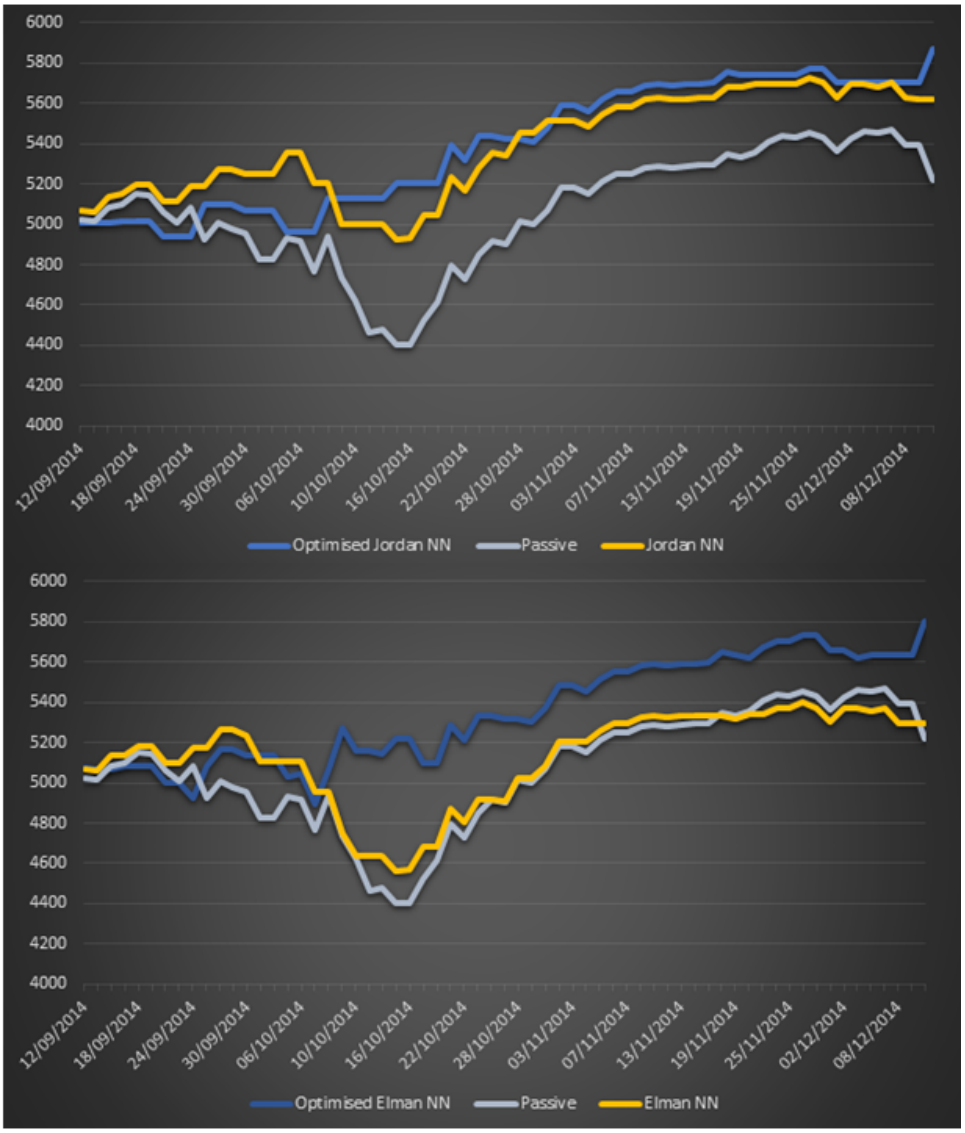

Figure 4.6.1: The three investment portfolios are presented on two separate charts, each related to Jordan Neural Networks (Jordan NN) at the top and Elman Neural Networks (Elman NN) at the bottom. The values on the y-axis denote portfolio values (£) with an initial value of £5,000. The trends in blue and yellow present the optimised models from the evolutionary optimisation algorithms and ordinary Neural Networks active investment portfolios respectively. The trend in gray represents the passive investment portfolio.

The active investment strategies use the input from the estimated Neural Network models and also technical analysis data variable K, called the Chaikin Oscillator, which determines the position of the forces of demand and supply - see details on the calculation of the variable in [63]. To maximise the investment portfolio, we employ an evolutionary optimisation algorithm. Given the objective investment function below:

$$\mathrm{f}(call, put) = \begin{cases} Invest_{n-1} + (Price_n - Price_{n-1}) & call \\ Invest_{n-1} + (Price_{n-1} - Price_n) & put \\ Invest_{n-1} & else \end{cases}$$

where
$call : Pred_n > a, \triangle K_{n-1} > b, \triangle K_{n-2} > c, \triangle K_{n-3} > d$,
$put : Pred_n < e, \triangle K_{n-1} < f, \triangle K_{n-2} < g, \triangle K_{n-3} < h$, $Pred_n$ is the predicted value at day n, $\triangle K_n$ is the change in Chaikin Oscillator at day n, and
a, b, c, d, e, f, g, h are variables whose values must be determined. In order to maximise the objective investment function, we consider the following maximization problem:

$$\begin{aligned} &\underset{a,b,c,d,e,f,g,h}{\text{maximise}} \quad \mathrm{f}(call, put) \\ &\text{subject to} \quad -0.4 <= b, c, d, f, g, h <= 0.4 \end{aligned} \tag{4.4}$$

The evolutionary optimisation algorithm is then applied to Equation (4.4) in order to generate the values for a, b, c, d, e, f, g, and h. The objective function is fairly optimised using just the first 35 days and the estimates obtained are kept constant to estimate portfolio values and trends for the next 100 days. Expectations regarding the relevance of this optimisation algorithms and technical analysis method are that trends obtained from the optimised models would be more stable than the ones that are not optimised. Also, we expect rising trends as these trends interpret to portfolio values. Decreasing trends would imply investment losses. Looking at the results from Figure 4.6.1 the optimised

active models outperform the ordinary estimated machine learning models and the passive portfolio. This conclusion is based on the fact that the trends in blue appear to be the most stable and fairly rising trends when compared with the trends from the ordinary estimated machine learning models. Even when persistent loss is reported in the passive portfolio in the period 07/10/2014 – 21/10/2014, trends from the optimised models appear fairly stable and rising. This is due to the fact that the optimised models take account of both bearish and bullish stock market using *put* and *call* options respectively.

4.7 Discussion and conclusion

This chapter delivers its first novelty by the nature of the data explored, which at our best knowledge, was not considered by previous studies. For analysis purposes, our framework combined the closing prices of S&P 500 constituents and their related sentiments which in total provides about 800 variables. This dimensionality challenge is n-fold increased due to lagging operation common with time series. To tackle the challenge of high dimensionality of the dataset in a computationally expensive prediction modelling approach that we proposed, a specially designed data pre-processing methodology was introduced. To the best of our knowledge, this is the first work to have used constituent sentiments and its closing stock prices (containing over 800 variables combining closing stock prices of the S&P 500 constituents and their respective sentiment data without lagging) in stock market predictive modelling.

With the rolling window of a 10-day predictions period and time-variant relationship between response variable and predictors - an approach which involves obtaining a new set of predictors for every rolling window - the analysis' challenge became compounded. The random forest method failed to do a good predictor selection, as a first method of choice that we considered. As such, we proposed a 3-step feature selection methodology involving the consecutive phases of variable clustering, PCA, and our own method of further feature selection that we call MBestGLM.

Having established the most significant variables in our proposed predictive modelling approach and justified the inclusion of sentiment in the approach as we proved its predictive value using Granger based methods, we develop models based on Recurrent Neural Network algorithms to predict the S&P 500 closing prices. However, this information per se is not sufficient to reliably predict the stock market trends and maximise investment portfolios. As such, we enhanced our approach by proposing investment strategy models which make use of the generated estimates from the predictive models as input variables to bridge these gaps. Results show that our proposed model appears to be stable even when the stock market is bearish and other approaches are failing. The rationale is that the proposed model is engineered to perform using *put* and *call* options during bearish and bullish moments, respectively. This represents another novelty of our work.

We currently develop further work on exploring the extension of this approach and of the approach proposed in our recent work [43], for several stock market indices.

In order to develop the computationally demanding approach that we proposed in this innovatory study, a parallel processing was performed using the R software package on a data analytics cluster of 11 servers with Xeon processors and 832GB of fast RAM. We note that approaches based on computationally cheaper predictive modelling techniques which encapsulate also feature selection, such as Lasso and Elastic Net, were investigated, among other several attempts in a preliminary phase of this study, and did not provide satisfactory results.

5

A two-step optimised BERT-based NLP algorithm for extracting sentiment from financial news

5.1 Motivation

Sentiment analysis involving the identification of sentiment polarities from textual data is a very popular area of research. Many research works that have explored and extracted sentiments from textual data such as financial news have been able to do so by employing Bidirectional Encoder Representations from Transformers (BERT) based algorithms in applications with high computational needs, and also by manually labelling sample data with help from financial experts. We propose an approach which makes possible the development of

quality Natural Language Processing (NLP) models without the need for high computing power, or for inputs from financial experts on labelling focused datasets for NLP model development. Our approach introduces a two-step optimised BERT-based NLP model for extracting sentiments from financial news. Our work shows that with little effort that involves manually labelling a small but relevant and focused sample of financial news, one could achieve a high performing and accurate multi-class NLP model on financial news.

5.2 Introduction

The internet is full of online expressions which could be in the form of social blogs, financial news, or other kinds of textual expressions - thanks to the advancement in computer systems. With this advancement comes the ease of accessing, storing and processing large amounts of textual data. And now, sentiment analysis which helps to detect the element of feelings in textual data, has become popular due to its vast applicability in areas such as artificial intelligence, stock market trading, politics, psychology, among others (Qie et al. [45], Bechara [2] and Hatfield [3]). For example, the polarity of sentiment extracted from textual data can be identified using sentiment analysis. This may be categorised as factual, positive reflecting a happy state of mind, negative referring to a sad mood, or neutral. In addition, one may also use sentiment analysis to assess the degrees of polarised sentiments by scoring the different polarities of sentiments.

Sentiment analysis in the domain of finance, especially where the sentiments obtained are used to improve the predictive power of stock market predictive models, is of utmost importance. The predictive value of sentiments is highly time-sensitive with respect to first mover advantage in the face of market imperfection (Vayanos and Wang [32]). That is, one would expect the prices of the stock market to reflect all available information. As new information becomes available, players in the market adjust their positions and this new information becomes fully incorporated into the prices. There is a gap between these two

points of information arrival and the time when prices reflect the new information. This short time window is termed as market imperfection (Vayanos and Wang [32]). This aspect supports the rationale behind the time sensitivity of the statistical significance of sentiment variables. Financial news could be the source of new information, expressed as sentiment, which has proved to be useful in enhancing stock market prediction with statistical and machine learning approaches (Smales [84], Shiller [114], Olaniyan et al. [110], and Marechal et al. [43]). Advanced NLP approaches are powerful tools that can be used to reliably and effectively extract sentiment polarity information from financial news, and we propose such a novel approach here based on adapting and extending the BERT algorithm [69].

Worryingly, it appears difficult - or so it seems - applying supervised NLP methods in this domain for two obvious reasons: 1) Developing NLP classification models requires a significant amount of effort to correctly label a huge amount of the training data to be used in the model training development. 2) The model to develop depends on the domain-specific corpus for learning transfer as opposed to any general corpora which are not well-suited for supervised tasks.

As a result of these concerns, NLP transfer learning methods have become a popular choice. They have been proven to be very promising and advanced the state of the art across natural language tasks. Moreover, the foundation of these models, the language model (LM) pre-training, is considered effective as the initial step required when developing natural language models (Dai and Le [8], Dolan and Brockett [131], Howard and Ruder [73], Baevski et al. [1]). The rationale for this choice of models is that they learn contextualized text representations by predicting words based on their contexts using very large corpora, and can be fine-tuned to adapt to downstream tasks (Peters et al. [92]). The challenge from the paucity of labelled data is avoided as the LM does not depend on it - rather, it predicts words from contexts based on the semantic information it has learnt. And the fine-tuning of the NLP transfer learning methods on labelled data uses the semantic information learnt to predict labels.

Here is where the problem lies: the fine-tuning of the NLP model on reliable labelled downstream tasks.

Manually labelling data for fine-tuning is a difficult task. First, it requires much time and effort. Second, the manually labelled data must be reliable and representative. Table 5.2.1 presents the experimental results on Financial PhraseBank [62] of 4845 financial news that were randomly selected from LexisNexis database and annotated by 16 financial experts. Interestingly, all the participants were able to agree on just 46% of the data's sentiment polarities. This clearly confirms the inherent challenge in manually labelling financial news data and, as a result, fine-tuning of models would suffer as a consequence.

Table 5.2.1: This table was taken from Araci [21]. Distribtution of sentiment labels and agreement levels in Financial PhraseBank

Agreement level	*Positive*	*Negative*	*Neutral*	*Count*
100%	%25.2	%13.4	%61.4	2262
75%-99%	%26.6	%9.8	%63.6	1191
66%-74%	%36.7	%12.3	%50.9	765
50%-65%	%31.1	%14.4	%54.5	627
All	%28.1	%12.4	%59.4	4845

Our framework is therefore centred on developing a reliable high-performing NLP model with no exposure to any of these aforementioned challenges.

The rest of this chapter is structured as follows: Section 5.3 introduces our proposed approach. Section 5.4 presents the basis of the NLP model that would be used in this work. Section 5.5 provides information about the data sources, the various datasets, and the methodology applied. Results from the use of the primary model are presented in Section 5.6. Section 5.7 details the rationales behind the use of the secondary model that we propose, and presents some empirical findings about why any NLP models with very high level of accuracy may perform poorly on real life data. Finally, section 5.8 provides the conclusion

to our work.

5.3 Proposed approach

Considering these challenges, our research focus is centred on developing a reliable NLP model for sentence polarity identification of financial news and we aim to achieve this with very minimal effort. In fact, our proposed two-step optimised BERT-based model overcomes these challenges. The main contributions to this work are therefore highlighted below:

1. We propose a two-step approach that includes: (a) a primary model that relies on the labelled data from the experiment results on Financial PhraseBank and an optimised BERT-based NLP model called the Roberta NLP introduced by Liu et al. [135], and (b) a secondary model that combines the experimental results and a small data sample of financial news data that has been manually labelled by us and validated with the primary model. This is to ensure that the secondary model has been fine-tuned with focused data.

2. We evaluate the primary model, and compare it with other related works in terms of their respective degrees of accuracy. The aim of this comparison is to see how the model fine-tuned with just the experimental results would perform on the financial news data related to the constituents of the S&P 500 index. The data is obtained from Intrinio platform [60].

3. We evaluate the results of the secondary model and compare with the primary model with the aim of assessing if the results obtained from both models are statistically different. Findings from the results would help us to understand the relevance of the secondary model, especially when it is trained on focused data. In addition, we aim to assess the quality of our proposed two-step BERT-based model that does not rely on high computing power and on inputs from financial experts on manually labelling focused data for fine-tuning.

We use the BERT model introduced by Devlin et al. [69] as the framework for developing our proposed two-step optimised NLP model. As the name implies, the centre-piece of the BERT model is the transformer which was first published by Vaswani et al. [11] and which is a great breakthrough in the world of language modelling. The section that follows will detail the BERT-based NLP framework.

5.4 BERT NLP

The likes of convolutional neural networks (CNN) and Long-Short Term Memory (LSTM) are useful language modelling, but there are some constraints around them. One of these constraints is their poor performance when it comes to processing long sentences - the probability of learning the contextual relations between words when they are far away from each other diminishes linearly (Kalchbrenner et al. [101]) or exponentially (Gehring et al. [70]) depending on the language model used. Although some transduction models -models that convert input sequence of elements into another output sequence - have been able to overcome this challenge through the coupling of neural nets with an attention learning mechanism that facilitates attention learning of specific words with the notion that these words could be embedded with contextual relevance (Bahdanau et al. [23], Kim et al. [134], Parikh et al. [9]). Another common problem with these transduction models is their inflexibility to parallel computation of tasks or their inefficient flexibility to it. This is where the transformer plays the leading role whereby it does not depend on any coupling of neural nets with attention mechanism. It uses its inherent self-attention mechanism solely to draw the contextual relations between input and output, and it also allows for efficient parallel computation of input and output (Devlin et al. [69]).

In the wake of the transformer, many language model pre-trainings have sprung up, and results from research works support the fact that these models are effective for enhancing NLP-related undertakings (Peters et al. [92], Howard and Ruder [73]). These models are applicable in a broad range of tasks such as

named entity recognition (Li et al. [74]), sentiment analysis(Sun et al. [19]), text summarisation (Miller [28]), among others.

Most of the pre-training-based models are unidirectional - from the left to the right - in learning the general language representations. Devlin et al. [69] state that such architectural constraint limits the choice of architectures in the first place and are sub-optimal for sentence-level tasks. In view of this BERT is proposed for fine-tuning because of its uniquely bidirectional approach for general language representations.

The optimised version of the pre-trained BERT model will be used in this work as originally presented by Liu et al. [135] for developing the primary model. This model will be the basis upon which our secondary model is developed.

5.5 Methodology

In the process of conducting sentiment analysis on financial news, we source for financial news data from Intrinio platform [60]. The data collected covers the period September 2012 and July 2019 and comprises 1.05 million records - our interest is to extract multiclass sentiment polarities from this data. The financial news data collected are related only to the constituents of the S &P 500 index. Our aim is to identify the sentiment polarities from this data by applying our proposed NLP model. Extra care is required in ensuring that false positives and false negatives are minimised. In doing this we propose a two-step optimised BERT-based NLP model where the first step is to produce the primary model that explores both the labelled data from the experimental results on Financial PhraseBank and an optimised BERT-based NLP model. The level of accuracy of the model would be examined to see if it qualifies enough to become our primary model. More specifically, we employ the Roberta NLP model which is considered the optimised version of the BERT model. We train the optimised BERT-based NLP model with the experimental results on the financial PhraseBank dataset which is the dataset used also in [107] and [21]. The dataset consists of 4,845 financial news that were randomly selected from the LexisNexis

database. In the process of manually labelling the financial news data, 16 financial professionals were asked to participate. 47% (2263 of the 4845) of the financial news had 100% agreements from all the participants. This implies that some sentences were assigned different labels by different participants. Clearly, this is a confirmation that manual labelling is complex and challenging to correctly assign true labels due to varyingly and subjectively contextual perspectives. It is also laborious to manually label a high volume of sentences for developing training models.

In view of these issues, we resolve to using only the financial news with 100% agreement level from all the participants totalling 2263 sentences in training our primary model. The summary of the selected sample data is presented in Table 5.5.1.

Table 5.5.1: Experimental results on the Financial PhraseBank dataset containing the 2263 financial news with 100% agreement level labelled by the 16 financial professionals.

Value	*Polarity*	Count
0	Neutral	1390
1	Positive	570
-1	Negative	303
Grand total		**2263**

The second step in our approach is to develop the secondary model that explores combined datasets from two data sources: a) the same dataset used in the primary model which is the experimental results from the Financial PhraseBank, and b) a small sample dataset of 2,000 records from [60] that has been manually labelled by us and also validated by the developed primary model. The aim is to see if by including focused labelled data the secondary model will outperform the primary model. Clearly, the primary model is a key and integral part of the secondary model. Attention would therefore be paid to the degree of

accuracy of the model with the assumption that a high degree would constitute its acceptance as a basis for developing the secondary model. Recall that 16 professionals were involved in the experimental labelled results on the Financial PhraseBank. One of the possible reasons for involving many financial professionals was to ensure that the experimental results produced were reliable and of high quality in the labelling task.

Understandably, going through at least the same level of effort of involving many financial experts in the manual labelling exercise is resource-consuming and time-taking. As a result, we are proposing the two-step NLP approach that we consider to be effective both in labelling and in model training. It is worth mentioning that the manual labelling of sentences itself is challenging not just in regard to the volume of sentences to label, but also in correctly labelling sentences because sometimes there seems to be a very thin line among the classes e.g. positive and neutral, for example.

Below are some examples:

1. InvestorPlace Stock Market News Stock Advice amp Trading Tips Apple NASDAQ AAPL will be reporting its third quarter earnings on July. Apple stock has performed well since the start of June posting a gain since June but on July all bets are off.

2. Why Apple Stock May Be a Case of Near-Term Pain, Long-Term Gain

3. UPDATE -Ireland invests disputed Apple taxes in low-risk bonds

4. American CEO reiterates confidence in Max return by mid-August despite unclear timetable from Boeing, FAA

5. UPDATE -Apple explores moving -% of production capacity from China - Nikkei

Manually labelling over 1 million financial news would be very laborious and we would expect a lot of disagreements in labelling some news among us, if we were to perform this task: hence, the need to manually label a small sample and

use the trained model to validate the results of our labelled news. Where we have disagreements in the results between our manual labelling and the trained model, we review carefully in order to identify the true labels. We are more interested in the false positives and negatives from the model's results so that we could review the sentences that we consider to be wrongly labelled and add the reviewed sentences to the training data and finally obtain the secondary model. Eventually, with help from the trained primary model, we have 2,000 labelled news - that have been randomly selected from [60] - to be added to the original training data. Our secondary model is therefore trained by combining the 2,000 labelled news items that have been reviewed and the experimental results on the financial PhraseBank. The trained secondary model is then applied to over 1 million financial news data in order to obtain sentence polarities which could be positive, neutral or negative.

5.6 Primary NLP model

Before the BERT NLP model could be used, it has to go through two key steps. First, the BERT would have to be pre-trained like every other language model, so that they could learn the contextual relations between words. Pre-training a model on a very large corpus is a very resource-consuming effort, especially with the amount of time and computing architecture capacity required. For example, most of the BERT pre-training exercises were conducted on the Google cloud([69], [139]) and Amazon cloud([21]) translating to the high dependence of the pre-training stage on high computing machines. Devlin et al. [69] pre-trained the model using the corpus that contained the combination of the BooksCorpus (800M words) (Zhu et al. [139]) and English Wikipedia (2,500M words).

The second stage would require that the pre-trained model goes through supervised learning where the training dataset contains texts and their respective labels e.g. "The US stock market is bullish" is the text and the label is "positive". The results predicted using the trained BERT-based models have been promising

and this accounts for its popularity.

Araci [21] examined if by both pre-training(unsupervised learning) and training (supervised learning) the BERT on downstream tasks could improve further the BERT model. In the process, the author pre-trained the BERT model on the financial news data obtained from Reuters at first, and then trained the model using the experimental results on the financial PhraseBank dataset which was the same data used in [107]. Findings showed that such process could improve the model performance by 15% in accuracy.

Liu et al. [135] revisited the work done by [69] and concluded that the pre-trained BERT was not at its optimal level. They pre-trained the model all over and finally obtained the optimised BERT model. In view of this development, we would be using the optimised BERT model to compare the results with the Araci [21]'s. Findings from this work would help us to answer the following questions:

1. Is pre-training the BERT model with targeted downstream task necessary as opposed to the huge corpus of BooksCorpus (800M words) and English Wikipedia used to pre-train the BERT?

2. Is the downstream data used by [107] and [21] for the supervised model training and evaluation representative enough of the financial news?

As shown in Table 5.6.1 the optimised BERT-based model appears to have achieved the same level of accuracy as the FinBERT model. The optimised BERT is only trained with the downstream tasks and the results show that it performs as highly accurate as the FinBERT's which was proposed by [21]. In view of this, it would be hard to conclude that pre-training the BERT model with downstream tasks would improve the accuracy level of the BERT as [21] has claimed.

The optimised BERT model performs well with a very high level of accuracy when trained on the Financial PhraseBank dataset. Would this trained model perform well on real life data? To answer the question, we apply the trained primary model to a new financial news dataset in order to evaluate its reliability when applied to a real life situation. This task is addressed in the next section.

Table 5.6.1: Most of the information in this table was taken from Araci [21] with reference to Malo et al. [107], Krishnamoorthy [120] and Maia et al.[95] regarding the results using LPS, HSC and FinSSLX respectively. The last row added by us represents the results obtained from the optimised BERT-based model - this report was based on a 5-fold cross-validation results.

Data with 100 % agreement			
Model	Loss	Accuracy	F1-Score
LSTM	0.57	0.81	0.74
LSTM with ELMo	0.50	0.84	0.77
ULMFit	0.20	0.93	0.91
LPS	-	0.79	0.80
HSC	-	0.83	0.86
FinSSLX	-	0.91	0.88
FinBERT	0.13	0.97	0.95
ROBERTA	**0.12**	**0.97**	**0.97**

5.7 Secondary NLP model

We use the experimental results on the Financial PhraseBank data to train the optimised BERT model and this achieves a high accuracy level of 97% as presented in Table 5.6.1. In order to assess how representative the training data is, we evaluate the predicted results obtained from the trained model (primary model) on a new set of financial news obtained from [60]. If the results obtained show at least 90% level of accuracy on the new data, we would conclude that the training data is highly representative of the financial news and that the model is reliable without the need for the proposed secondary model.

In this process we start by manually labelling 3000 financial news randomly sourced from [60]. We understand the concern that we might not be 100% accurate in the manual labelling; hence, the need to rely on the primary model for the validation of the manual labels and the review and correction of the labels

Table 5.7.1: This report represents the out-of-sample evaluation results from the primary model when applied to new focused data.

Class	Precision	Recall	F1-Score	Support
-1	0.93	0.93	0.93	42
0	0.86	0.59	0.70	63
1	0.77	0.93	0.84	91

Table 5.7.2: This report represents the out-of-sample confusion matrix obtained from the primary model when applied to new focused data. The confusion matrix leads to an accuracy of 82%.

		True Label			
	Label	-1	0	1	Total
Predicted Label	-1	39	1	2	42
	0	2	37	24	63
	1	1	5	85	91
	Total	42	43	111	196

that appear as false positives and false negatives.

The results from the model are presented in Tables 5.7.1 and 5.7.2. The results show that when the trained primary model is applied to predict the sentiments from a different sample data, the level of accuracy drops significantly to 82%. On this ground we conclude that the initial sample data - the experimental results on the Financial PhraseBank dataset - is short of being considered as a representative of the financial news in general. Considering this, we develop the secondary model which is the second step of our proposed two-step optimised BERT-based NLP model. The secondary model clearly shows improvement judging by the performances presented in Tables 6 and 7, including an overall accuracy of 99%. We should note that in repeated test experiments we obtained accuracies of at least 97%.

Table 5.7.3: This report represents the out-of-sample evalutation results from the secondary model when applied to new focused data.

Class	Precision	Recall	F1-Score	Support
-1	0.98	1.0	0.99	42
0	1.0	1.0	1.0	63
1	1.0	0.99	0.99	91

Table 5.7.4: This table represents the confusion matrix obtained from the secondary model when applied to new focused data. The confusion matrix leads to an accuracy of 99%

		True Label			
	Label	-1	0	1	Total
Predicted Label	-1	42	0	0	42
	0	0	63	0	63
	1	1	0	90	91
	Total	43	63	90	196

5.8 Discussion and conclusion

The application of BERT framework in NLP modelling has gained huge popularity, judging my its surprisingly high-performing and promising results in broad NLP-related endeavours. As a result of this outstanding success, many research works have embraced this framework. Some researchers have attempted to improve on the BERT-related work by proposing the pre-training of the BERT models on downstream tasks as opposed to the use of general corpora (Araci [21]). They claim that doing so would improve the accuracy level of the BERT-based models.

First, this suggested approach is clearly laboriously resource-consuming in that one would have to source for the focused corpus for the BERT pre-training. To further complicate the challenge, the pre-training would have to be performed on high computing machines. These are very daunting. Another clear concern is the

input required from professional experts on manually labelling the sample data for model fine-tuning.

We acknowledge how arduous manually labelling a high volume of financial news data for fine-tuning BERT-based models could be, the need to involve professional experts in the labelling effort, and the dependency on high computing machines in the pre-training process.

In view of these challenges, we propose a two-step optimised BERT-based approach which has the tendency of achieving sound results without the need for these 'requirements'. That is, with our proposed approach we consider it unnecessary to pre-train the BERT model. And we would need to manually label just a small sample for the fine-tuning stage. And noting that our manually labelled training data might contain some false labels our approach has therefore considered such occurrence by searching for false labels (otherwise known as mismatches between the results from the primary model in the first step and our manual labels) and correcting before they are fed into the secondary model which is the second step of our proposed model. This is done by establishing a primary model developed by first training the optimised BERT model using the experimental results from the Financial PhraseBank news data. The trained primary model is then applied to the small sample of 3000 manually labelled news for validation. In doing so, the false matches between the manual labelling and the results from the primary model are identified, reviewed and corrected accordingly. The manually labelled data that has been properly updated -reviewed and corrected - are then added to the initial training data resulting in a combined training data which would then be used to re-train an optimised BERT model. This becomes the trained secondary model. The rationale for these two steps is that the trained primary model on its own is not sufficient due to the misrepresentation of the initial training data as shown by its low accuracy level of 82%. But with the secondary model that has been developed with more focused data and little manual input, we are able to achieve a higher accuracy of over 97%. This is achieved with minimal and manageable effort.

This work can be further extended to explore the relationship between

financial sentiments and the stock markets. This can be achieved by using our model to extract the sentiment polarities from financial news.

6

Event-based algorithmic intraday trading

6.1 Motivation

"When" and "how" as simple as they look play key roles jointly in investment portfolio optimisations. Especially in the stock market, investors are keen to understand "when" and "how" to invest with the aim of profit-taking and/or stop-losing. Understanding these elements requires learning from historical information with respect to the stock market. Interestingly, intraday data for a year alone is voluminous and this can be challenging to explore. But more complexities are introduced where high frequency intraday stock market data is explored - such data contains both relevant and irrelevant information.

We explore a considerable volume of high frequency intraday data and introduce a predictive machine learning model that incorporates an event-driven algorithm. This event-driven algorithm helps to filter out the irrelevancies in the

data for model development. In addition, we propose a fractional differencing technique for achieving variable stationarity as opposed to using the popular method of log differencing with its inherent shortcoming of variable's memory loss and weakened predictive power.

Technical analysis indicators are also included as part of the model variables for predicting the directions of the stock market.

A machine learning framework such as support vector machine is used in developing our proposed stock market predictive model. Our aim is to demonstrate that with the right data and the appropriate selection of data processing techniques, our model can generate good and reliable results. In particular, the model developed can help with strategic investment decisions as to when and how to trade in the highly volatile stock market.

6.2 Introduction

A reliably good machine learning predictive model is expected to yield high financial gains while hedging against market losses. Models like this are hard to come by especially in the stock market world known to be full of many expectedly unknown surprises, market uncertainties, irregularities and frequent ups-and-downs volatilities (Marszalek and Burczynski [5],Abu–Mostafa and Atiya [122]). Yet, the searches for these good models continue.

First, the stock market is volatile. Exploring intraday financial data complicates the challenge of developing these stock market predictive models. Of course, it is not a problem training any machine learning predictive models on all the voluminous intraday data, but one should not be surprised if the outputs from these models are mere numbers with no human-friendly economic value (Lee [121]) - predictive models learn well when they are trained on relevant data (Lopez [105]). And with the growing popularity of high frequency trading that relies on intraday data, volatility modelling and event-driven sampling are now becoming popular among researchers. Abdersen and Bollersleve [128] were among the early researchers of the intraday volatility modelling. They developed

a framework for the integration of high frequency intraday data into the measurement, modelling and forecasting of daily low frequency return volatilities and return distribution. Their results showed that from just a simple Gaussian vector autoregressive volatility forecast incorporated with a parametric log-normal mixture distribution they got well-calibrated density forecasts of future returns.

The interesting work of [105] largely expanded on the capabilities of the intraday volatility modelling by incorporating a CUSUM (Cumulative Sum) technique initially introduced by Lam and Yam [81] to the volatility modelling. The basic idea behind the CUSUM technique supports the notion that data irrelevancies should not be fed into any predictive models. That is, only the data with informative elements should qualify to go into modelling. In this process, a data sampling technique that samples much more frequently during some noticeable events is proposed. And to identify these events, the volatility modelling is introduced. In our work we follow the same approach introduced by [105] in generating the sample required to train and test our model.

Some of the key questions about event-driven intraday sampling modelling are related to the positions to take for achieving profit and/or loss optimisations at any point in time. For simplicity, let us assume we are involved in an option contract which is an agreement between a buyer and a seller to trade an underlying security or index at a future date (Clarke et al. [112]). If we are buying a call option, we expect the price of the financial instrument to go up in the future so that profit can be made off the contract by exercising our right to buy the financial instrument. Clearly our positions on the contract today would be determined by what we think we know about the financial instrument now. That would require us giving much thought to some key questions such as: How do we know the future directions of the stock market of interest? Should we go long or short? Many investors and traders in at least the futures and option markets have found some comfort in the application of Bollinger Band Strategy (Ni and Zhang [98], Butler and Kazakov [89]).

In this work we use the meta-labelling technique introduced by Lopez [105] as

the basis for deciding on the market directions. We modify [105]'s work on the ground that we are interested in what happens immediately after there has been some market jumps or events in the stock market. This is one of the main aims of this chapter.

It is also worth mentioning that variable stationarity data pre-processing technique is a key element in developing a quality predictive model. As mentioned earlier, with the high volume of intraday financial data in collection, we intend to extract only the informative features for model development. This effort needs to be complemented with the right data preprocessing technique as well on account that the informative features extracted from the stock market data are highly likely to be non-stationary. Unfortunately, predictive models that consume stock market data require the inputs to be stationary. To this end, log differencing is a very popular data-preprocessing technique. But authors such as Alexander [14], Hosking [72] and Lopez [105] have voiced some concerns with the use of this technique. They argue that this technique causes variable (or data) memory losses leading to poor model performance. They therefore support an alternative approach known as the fractional differencing technique. When a variable is log differenced once, it implies that the variable is integrated at order one and this is denoted as $I(K = 1)$ where I denotes integration and K is the order for making a variable stationary. They argue that variables could also be made stationary for the value of $K < 1$. In this case the variables retain some memory insightful in the model development. Now, the unknown remains the value of $K < 1$ to use. We therefore propose a scalable algorithm for detecting the variables that are non-stationary, identifying the optimal values of $K < 1$ for making these variables stationary and processing them for stationarity. To the best of our knowledge, this is the first work to have developed such an algorithm without any need for human try-and-error approach as stated in [105].

The main contributions of this chapter are as follows:

1. We propose a stock market machine learning predictive model applicable to high frequency trading stock market with the aim of predicting the

directions of the stock market prices. The proposed model is incorporated with the event-driven CUSUM sampling introduced by ([81]), our proposed modified meta-labelling technique originally introduced by [105], and a set of technical analysis indicators.

2. We introduce a new technique for achieving variable stationarity as opposed to using log differencing for obtaining variable's integration of order $I(K < 1)$ in view of the fact that this long-preserved tradition diminishes the predictive power of variables. Instead, we propose a technique that identifies the non-stationary variables in a dataset and transforms these variables based on $I(K < 1)$ for achieving stationarity. We build on the fundamental approach of the fractional differencing originally introduced by Hosking [72].

3. We compare and evaluate the results from the two models built using the two datasets pre-processed based on the traditional log differencing and our proposed fractional differencing approaches respectively.

The rest of this chapter is organised as follows: Section 6.3 provides the information about the data sources explored. Section 6.4 details the stationarity techniques employed to our data. Section 6.5 presents our empirical findings from the use of both the log differencing and our proposed fractional differencing techniques and compare the results from both techniques. And finally, we have the conclusion in section 6.6.

6.3 Methodology

Two high frequency datasets are explored in developing our stock market predictive model. These datasets are purely stock-market-related. The main data source from where the datasets are created is the E-mini S&P 500 futures data which is obtained from the platform Kibot [53]. The data covers the period between April 2013 and July 2019 and it comprises over 1 million transactions.

The information is reported using the time zone in Chicago, Illinois, USA (GMT-5). At the initial stage, the data needs to be refined and properly structured before it can be used in modelling. To do this, two steps are followed. The first step is to sample the data so that only the relevant and informative observations and features are extracted for use in the model development. The other step involves the stationarity pre-processing of the relevant data. This effort ensures that the processed data satisfies some statistical and modelling properties like stationarity.

6.3.1 Sampling methodology

The E-mini S&P 500 futures information is provided in the form of tick data/bars meaning that the data containing bid, ask, size and price are recorded and updated whenever a trade occurs and they represent the "national best bid and offer" (NBBO) prices across multiple exchanges and electronic communication networks (ECNs), and they are not reported in any aggregated form according to [53]. In other words, the data obtained is unstructured. To make it meaningful and useful, the data needs to be properly structured. The most popular option is that it is structured in the form of time bars by sampling the data at a constant time interval, e.g. once every second, minute, daily or monthly. Sadly, this approach has some drawbacks, and these could not be simply overlooked. In real life, financial information is not processed at a constant time interval. More specifically, we expect low and high activities in the stock markets - the high activity period during the opening period for example. Oversampling therefore occurs during the low activity period and vice versa, as the difference in the activity periods is not considered. It has also been found that time-sampled data exhibits poor statistical properties e.g. non-normality of returns, heteroscedasticity, among others (Easley et al. [24]).

This section covers how the data obtained in its original form is structured and the sampling method employed in extracting the informative features. Highlighted below are the processes followed:

1. Data structuring - it details the approach employed in having the data structured.

2. Event-driven sampling - we endeavour to filter the irrelevancies from our data knowing that models learn well from the informative observations. To this end, we employ an event-driven sampling method to filter for the relevant observations.

3. Strategic trading positions and observation labelling - in investment trading it is important to identify strategic trading positions. By doing this we manage our profit-taking and/or loss limiting positions.

Data structuring

According to Investopedia [52] E-mini is considered as an electronically traded futures contract that represents a fraction of the value of its corresponding standard futures contract. E-mini future contracts are predominantly traded on the Chicago Mercantile Exchange (CME) and are available on a wide range of indices such as the NASDAQ 100, S &P 500, S &P MidCap 400, and Russell 2000, commodities such as gold, and currencies such as the euro. These contracts specify the cash flows to be exchanged among the holders and writers of the contracts whose values rely on the underlying S &P 500 Indices (Harris [86]).

We use the E-mini S &P 500 futures in this chapter. The data contains the bid price (bid) which is the highest price a buyer is willing to pay, the ask price (ask) known as the lowest price a seller is willing to sell, the price (price) as the price paid for the contract(s), and the size (size) representing the volume of transactions/contracts recorded at the specified price at a given time. The stock market and technical analysis indicators are extracted from the E-mini S &P 500 futures. In total, 1.8 million transactions are collected.

As mentioned earlier, the raw data is unstructured. So, we apply the dollar bar approach to refining it. Let us look at some examples that explain more clearly the methodology behind the data sampling and structuring.

Given that we have 10 transactions with the following time stamps $t_1, \ldots, t_{10}$ with the condition that the time stamps are not equally spaced but occurrences are sequential. That is, $t_n - t_{n-1} \neq t_{n-1} - t_{n-2}$ where $n \in [1, 10]$ and $t_n > t_{n-1}$. Corresponding to these time stamps we have the following prices $[100, 110, 120, 110, 110, 130, 120, 150, 100, 100]$ and sizes $[1000, 1000, 1200, 1000, 800, 900, 1000, 1000, 1500, 1200]$. This is better illustrated in Table 6.3.1.

Table 6.3.1: Tabular illustration of Dollar bar sampling.

Time stamp	Price	Size	Dollar value	Cumulative sum
t_1	100	1000	100,000	100,000
t_2	110	1000	110,000	210,000
t_3	120	1200	144,000	**354,000**
t_4	110	1000	110,000	110,000
t_5	110	800	88,000	198,000
t_6	130	900	117,000	**315,000**
t_7	120	1000	120,000	120,000
t_8	150	1000	150,000	270,000
t_9	100	1500	150,000	**420,000**
t_{10}	100	1200	120,000	120,000

Let us assume that the pre-determined value for dollar bar sampling is 300, 000 and that the dollar value is defined as $P_t * size_t$ where P_t is the stock price and $size_t$ is the stock size. For sampling we start by having the cumulative sum of the dollar value until we get the cumulative sum that is at least 300, 000. From our example we can see that the first one is achieved at the time stamp t_3 because the corresponding cumulative sum of the dollar value of 354, 000 is greater than the threshold of 300, 000 which implies that it qualifies into our sample data. We then restart the cumulative sum from the next dollar value after t_3. We can see that the next time stamp to qualify is t_6. This goes on and on and the results obtained become the sample data. The values in bold text under the

Cumulative sum column in Table 6.3.1 indicate the values to qualify into our sample data. This is therefore the approach followed in sampling the E-mini S &P 500 futures data as opposed to sampling at a constant interval.

In light of these concerns, we use the dollar bar sampling approach proposed in [105] whereby sampling takes place every time a predefined market value is exchanged. In this case, the number of shares is a function of the dollar value exchanged. This approach is considered robust to high price variations [105] and the drawbacks highlighted in the time-sampled data are addressed in this robust but yet unpopular approach.

Event-driven sampling

The previous subsection has detailed how the dollar bar has been applied to structure our raw stock market data sampled per second of time. Unfortunately, the structured data in its state is still not good enough to have it fed into any machine learning algorithms on the ground that the data contains a lot of irrelevant features and it would be good to have them filtered out. Generally, predictive models perform well when they are trained on informative features. But, in our case we have over 1 million transaction records. The challenge is how to extract only the informative features from the data so that our predictive model could be trained on the selected informative features. One common approach is to apply an event-based sampling approach to generate the relevant training data.

Easley et al. [25] revealed the potential problem from the use of time-sampled data. They argued that today's markets are being operated by automated algorithms with little human intervention and this may result in information oversampling during low-activity periods and vice versa. This implies that the stock markets are more active and process more information during the opening hours than during the period around noon or night, depending on the products. But automated algorithms trained on time-sampled data disregard the time sensitivity of activities.

Astrom et al. [80] examined an event-based sampling using a simple first order

system with event-based sampling and compared the achieved closed loop variance and sampling rate with results from their time-sampled data. The paper concluded that event-based sampling performs better than time-based sampling.

Event-sampling techniques require the observation of activities and the detection of regime changes or structural breaks leading to deviations away from the mean of a time series e.g. spikes in volatilities. One such technique that can help to detect these changes is known as the Cumulative Sum (CUSUM) technique developed in the fifties (Page [39], Page [40], Kemp [82] and Kemp [83]). The CUSUM technique is designed to detect a shift in the mean value of a measured quantity away from a target value. Consider IID observations $y_1, y_2, y_3, \ldots, y_t$ representing a stationary process. Let us assume that we are interested in detecting upward deviations. We define the cumulative sums as:

$$C_t = max\{0, C_{t-1} + y_t - E_{t-1}[y_t]\} \tag{6.1}$$

with boundary condition $C_0 = 0$. A CUSUM procedure would activate action at the first satisfying $C_t \geq h$ for a threshold value h. Extending the CUSUM procedure to cover downward deviations we have the CUSUM filter defined as:

$$\begin{aligned} C_t^+ &= max\{0, C_{t-1}^+ + y_t - E_{t-1}[y_t]\}, C_0^+ = 0 \\ C_t^- &= max\{0, C_{t-1}^- + y_t - E_{t-1}[y_t]\}, C_0^- = 0 \\ C_t &= max(C_t^+, -C_t^-) \end{aligned} \tag{6.2}$$

Eq. (6.2) is applied to sample our data whenever the condition $C_t \geq h$ is satisfied, at which point C_t is reset. This CUSUM technique is applied to our structured data.

Many research works do not support the use of time-sampled data. We move away from it as well and rather support the exploitation of an alternative approach which is an event-driven sampling approach. In this case we do not assume any

constant time intervals between the estimation data points. Therefore, in this process of extracting our event-driven data we begin with volatility estimation.

The stock market is influenced by many events. Each event in isolation could result in structural breaks in the stock market. For example, Huang [106] and Lobo [13] studied the impacts of political risk element and election on the stock returns and volatility. Regarding the stock market, it was revealed that the stock returns were negative in the election year and positive in the preceding years. It was also found that the stock volatility was very high during the period. These findings led to the conclusion that political events such as elections create uncertainty as a political risk factor for the stock market and that this factor influences the stock market. Stock market volatility often ignites stock market uncertainties. In view of this, we attempt to sample the stock market index of interest more frequently during some events in the market using the stock market volatility as a point of activation. This is calculated using the daily volatility at the intraday estimation points applied to a span of 100 days to an exponentially weighted moving (EWM) standard deviation. The Python function for estimating the volatility is stated in SNIPPET 3.1 of [105].

Having estimated the volatility, we now apply the CUSUM technique expressed in Eq. (6.2) to extract our event-driven sample but with some modification. Eq. (6.2) assumes that the value h is a constant threshold value. In our case, we assume the threshold value to be time-dependent and redefine it as $h_t = v_t * \theta$ where v_t is the volatility-based EWM standard deviation and $\theta = 0.3$.

With these significant elements of information, we proceed to determining our trading positions which are detailed in the next subsection.

Strategic trading positions and observation labelling

One of the main purposes of our research work is to be able to develop reliable predictive models for predicting the directions of the stock market. The scope does not cover predicting by how much a stock market price would change under any event scenario. Rather, we focus on the strategic positions we should take for

any investment actions. Three strategic trading positions would be considered. The first position would be to go long - buying a stock in the hope that the stock price would go up in the near future. The second position would be that we go short - selling a stock in the hope that the price would drop and then re-buy when it does. And then we have the hold position implying that we retain our current position without going long or short. Consider a feature matrix X with M rows, $X_{j=1,2,\ldots,M}$. For an observation X_j it would be associated with a label $y_j \in [-1, 0, 1]$ and the condition upon which the label value would be assigned is defined below:

$$y_t = \begin{cases} -1, & \text{if } r_{t_{i,0},t_{i,0}+\gamma} < -\tau \\ 0, & \text{if } \mid r_{t_{i,0},t_{i,0}+\gamma} \mid \leq \tau \\ 1, & \text{if } r_{t_{i,0},t_{i,0}+\gamma} > \tau \end{cases} \tag{6.3}$$

where the stock price return $r_{t_{i,0},t_{i,0}+\gamma}$ is defined as $r_{t_{i,0},t_{i,0}+\gamma} = \dfrac{p_{t_{i,0}+\gamma}}{p_{t_{i,0}}} - 1$, γ is a constant time step between the time $t_{i,0}$ and $t_{i,0} + \gamma$ when considering time-sampled data, and τ is a constant pre-defined threshold value representing the size of a return that determines the trading position one should take.

With help from the CUSUM technique in the previous subsection we have been able to extract event-driven data. But in trading one must take into consideration other cost factors e.g. trading commission. As a result, it is insightful to trade only when the benefits from doing so outweigh the costs. That is the basic rationale behind the threshold constant τ.

Recall that Eq. (6.3) is established on the assumption of constant time step γ implying that the equation is only applicable to a constant time series. Therefore, the stock market return $r_{t_{i,0},t_{i,0}+\gamma}$ is associated to time-sampled data. [105] modified this original equation to also cover event-driven sampling. As we are also exploring an event-driven sampling we therefore follow the same direction as [105]'s but with some proposed modification. First, we take into consideration the complete investment trading time horizon. That is, the stock market price $p_{t_{i,0}}$ at the first time $t_{i,0}$ that one secures a trading position and the price $p_{t_{i,0}+\gamma}$ at the time $t_{i,0} + \gamma$ that one comes out of trading.

For the purpose of identifying the time to enter a trading position and time to leave we apply the SNIPPET 3.6 in [105] and for generating the label values y_t we modify the SNIPPET 3.7 in [105]. Our modified version presented in Appendix A.0.1 assumes that the label y_t in Eq. (6.3) considers prices immediately after events have occurred as opposed to using the prices at the points where events occur. This will normally be achieved through automated trading algorithm development.

6.3.2 Stock market variables

Table 6.3.2 presents the variables explored as stock market indicators. These variables are derived from the S&P 500 E-mini. Spread is the difference between the ask and bid prices. Standard deviations with different rolling windows of 13, 30 and 50 are used to denote the price volatility variables.

Table 6.3.2: Stock market predictive indicators.

Index	Variable name
1	Stock price
2	Lags(1-5) of stock price
3	Average price with a rolling window of 20 (price_avg)
4	Lags(1 - 5) of rolling average price
5	Price volatility with a rolling window of 50
6	Price volatility with a rolling window of 30
7	Price volatility with a rolling window of 15
8	Spread
9	Size

6.3.3 Technical analysis variables

There are many variables that could qualify as technical analysis indicators, but we use 9 of them as presented in Table 6.3.3. From a technical analysis

perspective stock prices are expected to be contained within the upper and the lower bands. It implies that prices would normally operate within the bands, of course, depending on the value used to multiply the standard deviation. The functional relationships among standard deviations, the upper and lower bands can be expressed as:

$$\begin{aligned} upper &= RollingMean + \rho * std \\ lower &= RollingMean - \rho * std \end{aligned} \tag{6.4}$$

where *upper* is the upper band, *lower* is the lower band, $\rho \in (0, 4]$, *std* is the standard deviation, and *RollingMean* is the rolling mean. ρ normally determines the percentage of the stock prices contained within the upper and lower band. The higher the value, the more contained are the prices in the bands. In our case we use 1.5, and for the rolling mean the window of 50 is arbitrarily selected.

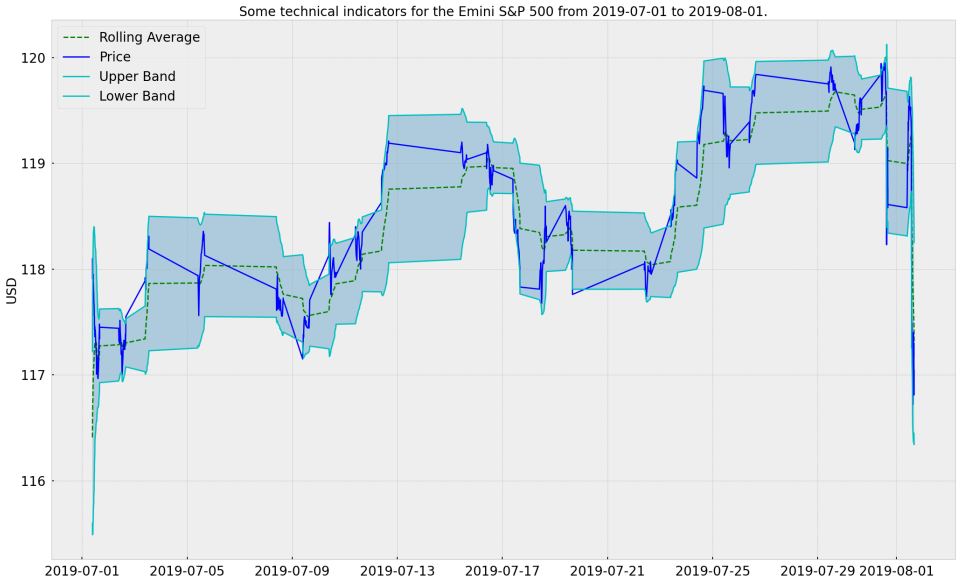

Figure 6.3.1: The figure depicts the relationships among price, average price, upper and lower bands. ρ is assigned with a value of 1.5.

As displayed in Figure 6.3.1 most of the stock prices are contained in the bands. There is a popular perception among technical analysis traders that when the stock price touches the upper band, then a fall in price would be expected and vice versa ([98], and [89]). The pivot point variable is used to determine the overall market strength. When the stock price cuts through above it, the market is considered to be bullish, and vice versa.

Table 6.3.3: Technical analysis predictive indicators.

Number	Variable name
1	Pivot point
2	Supports (S1,S2,S3)
3	Resistances (R1,R2,R3)
4	Chaikin oscillator (Chaikin)
5	Relative strength index (RSI)
6	Serial autocorrelation with a rolling window of 50
7	Lags (1-5) of serial autocorrelation (price_autocorr_1,...,price_autocorr_5)
8	high (maximum value within a rolling window of 20)
9	low (minimum value within a rolling window of 20)
10	Upper band (upper)
11	price_uppr (price - upper)
12	Lower band (lower)
13	price_lwr (price - lower)

Resistance, or a resistance level, is the level where there is pressure on stock prices to come down. Every seller would like to sell when the price of an asset is high. As the stock prices move up, there is a growing number of sellers willing to sell. As the prices reach this level, there is an expectation of a price trend reversal. This notion may not hold true especially when there is new information that could disrupt market attitudes. New information could have a positive impact on the asset and the trend shoots upwards beyond the resistance. Support is related to falling prices. When prices keep falling, it leads to a growing number of buyers

bidding for stocks. This action adds pressure to the stock prices and finally results in trend reversal ceteris paribus. Figure 6.3.2 presents the relationships among resistance, support, price and pivot point. It shows how the stock price cuts above the pivot point, for example, confirming the expected upward price trends and vice versa.

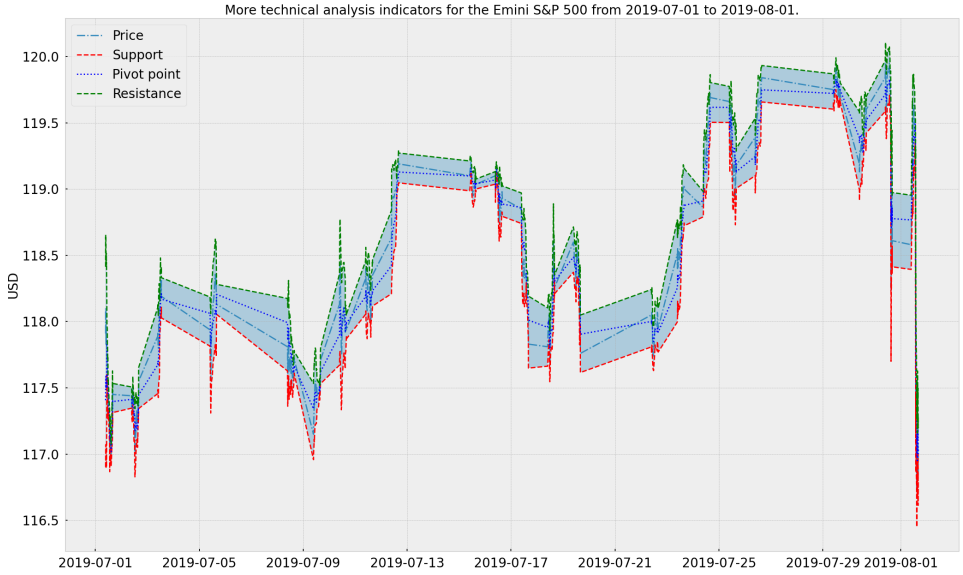

Figure 6.3.2: The figure depicts the relationships among pivot point, price, resistance and support.

To obtain the pivot point (PP), resistance (R_1,R_2,R_3), and support (S_1,S_2,S_3) variables, we use the high and low stock prices. In our work we use maximum (high) and minimum (low) prices with a rolling window of 20. Then

we define the technical analysis indicators as:

$$\begin{aligned} PP &= (high + low + price)/3 \\ R1 &= 2 * PP - low \\ S1 &= 2 * PP - high \\ R2 &= PP + high - low \\ S2 &= PP - high + low \\ R3 &= high + 2 * (PP - low) \\ S3 &= low - 2 * (high - PP) \end{aligned} \tag{6.5}$$

The concept of the relative strength index (RSI) was first proposed by Wilder [75]. It is considered a momentum indicator that assesses the magnitude of recent price changes. This assessment is used to determine if a stock is being overbought or oversold and can have a value between 0 and 100. When the RSI is above 70, the stock in the market would be considered as being overbought and this may add some corrective pressure for a trend reversal. When the RSI is below 30, it would mean that the market is in an oversold situation.

In the process of calculating the RSI the first step is to estimate the magnitude of the price changes. Recall that our *high* and *low* variables are based on the maximum and minimum prices with a rolling window of 20 respectively. So, these are used to first estimate the magnitude of the price changes as expressed below:

$$\mathrm{U}pward_t = \begin{cases} high_t > high_{t-1} & high_t - high_{t-1} \\ else & 0 \end{cases}$$

$$\mathrm{D}ownward_t = \begin{cases} low_{t-1} > low_t & low_{t-1} - low_t \\ else & 0 \end{cases}$$

where t is the time point. Next step is to estimate the RSI which is the ratio of the

n-period exponential moving averages (EMA) of $Upward_t$ and $Downward_t$ respectively. A 14-period EMA is very popular, and we would be using the same value. RSI is finally obtained using Eq. (6.6).

$$\begin{aligned} RS &= EMA(Upward, n)/EMA(Downward, n) \\ RSI &= 100 - 100/(1 + RS) \end{aligned} \tag{6.6}$$

where n represents the 14-period.

The Chaikin oscillator (Chaikin) is also a popular technical indicator that measures market movement. This is established by estimating the accumulation-distribution line of the moving average convergence-divergence using the difference between the 10-period EMA of the accumulation-distribution line and the 3-period EMA of the accumulation-distribution line. The equation is expressed below:

$$\begin{aligned} ad &= (2 * price - high - low)/(high - low) * size \\ Chaikin &= EMA(ad, 3) - EMA(ad, 10) \end{aligned} \tag{6.7}$$

6.4 Stationarity

In statistical analysis, one of the most basic requirements is the ability of the financial data under consideration to exhibit some statistical properties of invariance over time. This statistical element relates to the stationarity hypothesis defined as:
Given a set of time instants $t_1, \ldots, t_n$ and time interval τ the joint distribution of the returns $r(t_i, T), \ldots, r(t_n, T)$ is the same as the joint distribution of the returns $r(t_{i+\tau}, T), \ldots, r(t_{n+\tau}, T)$ and since τ does not affect $r(\cdot)$ then r is not a function of time. As most financial data are known to be non-stationary differencing at first order is very common in virtually all the financial time series literature. But why is differencing at first order considered as the optimal level? It seems the

preference is based on the popularity of log differencing - and it is simple to apply.

6.4.1 Log differencing for stationarity

We examine the variables at the level in order to see if they are stationary or not. In doing so we apply the Augmented Dickey Fuller stationarity test on the variables and as expected, most of the variables are not stationary. For example, the S&P 500 E-mini stock price is not stationary at level. To make them stationary the prices are log differenced. Model performance is assessed based on this traditional stationarity pre-processing technique.

6.4.2 Proposed fractional differencing technique for stationarity

Alexander [14] and [105] expressed some concern in the use of first order differencing. The papers mentioned that the use of first order differencing makes data transformation stationary but at the expense of memory loss. They propose the idea of fractional differencing which still achieves stationarity and, at the same time, keeps the relevant signal or memory that improves a model's predictive power. Interestingly, this technique was mentioned first in Hosking [72] but it still remains almost non-existent in recent literature. The general overview of the technique can be expressed in the form of a difference operator as presented below.

Given the backshift operator B and a matrix of real-valued features $\{X_t\}$ with the relationship between them expressed as $B^k X_t = X_{t-k}$ for any integer $k \geq 0$. For example, $(1 + (-B))^{n=2} X_t \equiv (1 + 2(-1)B + (-1)^2 B^2) X_t$ becomes $(1 - 2B + B^2) X_t = X_t - 2X_{t-1} + X_{t-2}$. For a real number f, $(1 + (-B))^f = \sum_{k=0}^{\infty} \binom{f}{k} (-B)^k$ as a binomial series. And with respect to

fractional model its binomial series expansion could be expressed as:

$$
\begin{aligned}
(1+(-B))^{f} = \sum_{k=0}^{\infty} \binom{f}{k} (-B)^{k} &= \sum_{k=0}^{\infty} \frac{\prod_{i=0}^{k-1}(f-i)}{k!}(-B)^{k} \\
&= \sum_{k=0}^{\infty} (-B)^{k} \prod_{i=0}^{k-1} \frac{f-i}{k-i} \qquad (6.8) \\
&= 1 - fB + \frac{f(f-1)}{2!}B^{2} - \frac{f(f-1)(f-2)}{3!}B^{3} + \ldots
\end{aligned}
$$

Finally, when the binomial series expansion in Eq. (6.8) is applied to the matrix of real-valued features X_t the results would become

$$
(1+(-B))^{f} X_t = (1 - fB + \frac{f(f-1)}{2!}B^{2} - \frac{f(f-1)(f-2)}{3!}B^{3} + \ldots)X_t \qquad (6.9)
$$

First order differencing of time series at level, as explained in the previous subsection 6.4.1 is equivalent to Eq. (6.9) for $f = 1$ as demonstrated below:

$$
\begin{aligned}
(1 - B + \frac{1(1-1)}{2!}B^{2} - \frac{1(1-1)(1-2)}{3!}B^{3} + \ldots)X_t &= (1 - B + 0 - 0 + \ldots)X_t \\
&= (1-B)X_t \\
&= X_t - BX_t \\
&= X_t - X_{t-1} \\
&\qquad (6.10)
\end{aligned}
$$

A second order differencing is also equivalent to Eq. (6.9) for $f = 2$ as

presented below:

$$
\begin{aligned}
(1 - 2B + \frac{2(2-1)}{2!}B^2 - \frac{2(2-1)(2-2)}{3!}B^3 + \ldots)X_t &= (1 - 2B + B^2 - 0 + \ldots)X_t \\
&= (1 - 2B + B^2)X_t \\
&= X_t - 2BX_t + B^2X_t \\
&= X_t - 2X_{t-1} + X_{t-2} \equiv (X_t - X_{t-1}) - (X_{t-1} - X_{t-2})
\end{aligned} \tag{6.11}
$$

In almost all cases we expect time series to be stationary at order $I(1)$ when they are not stationary at order $I(0)$, in its original form. The idea of fractional differencing is that time series that are not stationary at $I(0)$ can still be transformed to become stationary at order $I(0 < f < 1)$ instead of at order $I(f = 1)$. We therefore propose a novel approach built upon the fractional differencing technique developed by [105]. Our approach automatically examines each variable in a data for stationarity and transforms it to a stationary series where necessary based on the individual optimal value of f for each variable. Optimal value is the least value between 0 and 1 of the parameter f that makes a time series variable achieve stationarity.

6.5 Stock market predictive models

In this section we apply a simple machine learning framework to two different datasets - one that has been stationarised using log differencing and the other that has been processed using our proposed optimal fractional differencing technique. These two datasets are denoted as order of integrations $I(f=1)$ and $I(f<1)$ respectively. We are keen to understand how the model developed with the dataset of $I(f<1)$ would compare with the model developed with the data of $I(f=1)$. To this end we split each of the datasets into three subsets. The first is the training data that we employ in training the model. Then we have the test data to evaluate the model performance. Finally, there is the validation data for

validating the model performance, consistency, reliability, and accuracy.

The fact that the intraday financial data is the basis for our model does not necessarily imply that we are working with a high volume of data on the account that we are only interested in the event-driven sampled data. That is, we explore only the sample data that has been obtained using Eq. (6.2). This sampled data requires some filtering as we are interested in the sampled data that satisfies our investment return threshold as defined in Eq. (6.3) where the minimum expected return τ is assigned the value 0.2 .

The processed stock market data contains two labels suggesting when and how to go short or long. We refer to the holding period during the labelling, but this position would not be included in the model development as part of the labels, but it would still remain an integral part of the model. For example, the model we are developing would suggest when to enter a trading position, how long to hold the position, and when to exit the trading position. Referring to Eq. (6.3), $t_{i,0}$ is denoted as the time to enter the trading position, γ is the holding period and the time to exit is represented as $t_{i,0} + \gamma$.

In total, we have 2,035 trading positions which represent the data points to explore in developing our predictive models. Table 6.5.1 presents how the data is split into training, test and validated datasets.

Table 6.5.1: Datasets for model development and validation.

	Datasets			
Label	Training	Test	Validation	Total
-1	491	164	164	819
1	730	243	243	1216
Total	1221	407	407	2035

We explore the support vector machine (SVM) framework for simplicity. We evaluate the model quality and performance and also compare the results on the two datasets processed by using synthetic minority oversampling technique (SMOTE) and stratified 10 folds to achieve balanced classification between the

labels as shown in Table 6.5.2.

Table 6.5.2: Datasets processed by SMOTE to achieve a balanced training dataset.

	Set: Training dataset balanced		
Label	Training	Test	Validation
-1	730	164	164
1	730	243	243
Total	1460	407	407

In the next subsections we would present the results obtained from applying our trained SVM-based predictive models on both the log differenced and fractionally differenced datasets. Findings would help us to determine if log differencing for stationarity reduces the predictive power of variables and if the proposed fractionally differencing technique is more effective than log differencing technique.

Our model trained on log differenced dataset

The primary aim in this subsection is to evaluate the effectiveness of our predictive models in correctly predicting the directions of the S&P 500 E-mini options using log differenced stationary dataset (I(1)). Presented in Tables 6.5.3 and 6.5.4 are the results from our trained model. Based on the recall, precision, accuracy, and confusion matrix of the model performance on the test dataset, we can agree that our model is able to predict the directions of the stock price. The trained model is also employed on the validation dataset and the results obtained once again confirm the high accuracy of our model. In view of these findings, we conclude that our variables of interest would be able to capture the directions of the stock prices.

In light of these findings, we would also like to see if the fractionally

Table 6.5.3: Support vector classification evaluation results from log differenced (I(1)) test and validation datasets.

	Training dataset balanced			
Label	Precision	Recall	F1-Score	Support
-1	0.56	0.59	0.57	163
1	0.71	0.70	0.70	243
Test sample accuracy		0.65		
Validation accuracy		0.67		

Table 6.5.4: Support vector classification confusion matrix results from log differenced (I(1)) test dataset.

		True Label		
	Label	-1	1	Total
Predicted Label	-1	96	68	164
	1	74	169	243
	Total	170	237	407

differenced dataset would provide at least the same level of performance.

Our model trained on proposed fractionally differenced dataset

As we have seen from the previous subsection the model developed using the I(1) dataset captures well the directions of the stock prices. We are keen to also explore if our proposed SVM-based predictive model trained on the fractionally differenced dataset would provide at least the same level of performance. The works from [72], [14] and [105] have shown that fractionally differenced data retain some memory that helps to improve the predictive power of variables. This subsection would therefore be tasked with two key aims: examine if the model results from using I(<1) would be able to predict the directions of the stock prices; and if the trained model on I(<1) dataset would outperform the trained model on I(1) dataset judging by the values from precisions, recalls, F1-Score and accuracies. Tables 6.5.5 and 6.5.6 display the results from the model trained

on I(<1) dataset.

Table 6.5.5: Support vector classification evaluation results from fractionally differenced (I(<1)) test and validation datasets.

	Balanced training dataset			
Label	Precision	Recall	F1-Score	Support
-1	0.61	0.65	0.63	163
1	0.75	0.72	0.74	243
Test sample accuracy		0.69		
Validation accuracy		0.70		

Table 6.5.6: Support vector classification confusion matrix results from fractionally differenced (I(<1)) test dataset.

		True Label		
	Label	-1	1	Total
Predicted Label	-1	106	58	164
	1	68	175	243
	Total	174	233	407

As we can see the values of the precisions, recalls, F1-Score and accuracies from Table 6.5.5 are all higher respectively than the ones from Table 6.5.3 clearly confirming the findings from [72], [14] and [105] - log differencing erases memory from data. In light of this we conclude that fractionally differencing data for stationarity is more effective and retains the relevant memory that helps to improve the predictive power of variables. The final list of variables used in these models are also presented in Figure 6.5.1 where *autocorr_4* for example is the 4th lagged price serial autocorrelation. From the symbol *price_frac_t3*, *price* denotes the stock price, *_frac* implies that the variable has been fractionally differenced, and *_t3* indicates the 3rd lagged price. Most of the technical analysis indicators appear to be relevant.

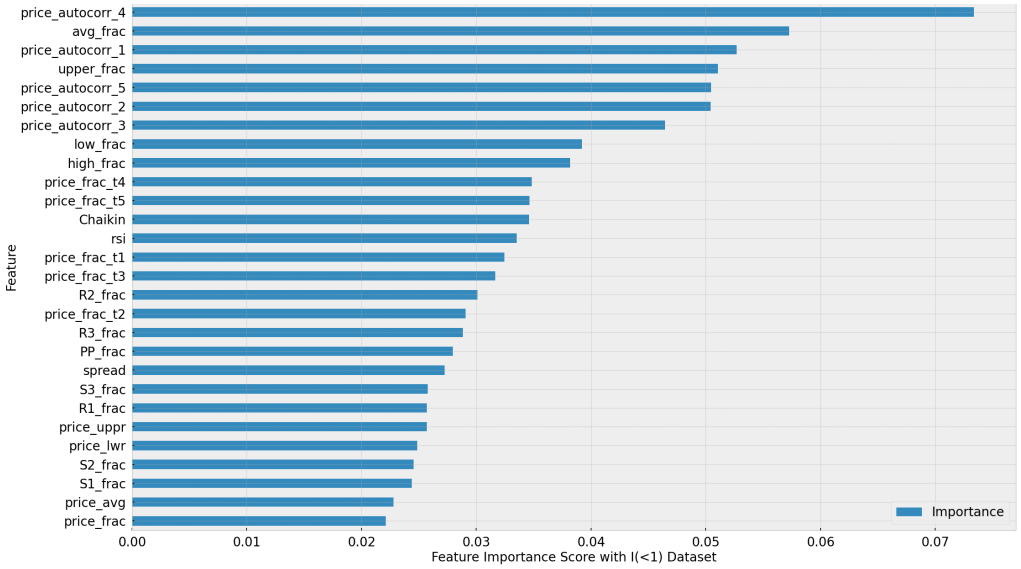

Figure 6.5.1: The variable importance representation of the selected variables.

6.6 Discussion and conclusion

We explore our proposed fractionally differenced E-mini S&P 500 intraday data to predict the directions of the stock prices by developing a stock market predictive model. In this process of model development, we employ the CUSUM technique for data sampling. On account of the high volume of intraday stock data, the CUSUM technique is crucial for extracting only the informative features from the data. So, the CUSUM technique is employed to sample more frequently during some events. To determine events within the stock market, we make use of a time-varying return volatility.

For identifying strategic trading positions regarding when to enter and exit trading positions, we modify the approach proposed by [105]. We would expect that trading positions are taken immediately after some events have been identified. This is therefore one of the modifications introduced in [105]'s work. This is one of the contributions to this work.

Another notable contribution is the introduction of a stationarity algorithm

that detects variables with white noise in a dataset, optimally stationalises them with an order that is less than 1. For example, when a variable is long differenced, it is denoted as I(1) implying that a series is stationary at order one. [72] expresses some concerns that variables at I(1) may have been over-differenced and this may cause memory loss which could adversely impact the predictive power of the variables. As a result, [72] suggests that variables should be allowed to retain some memory as long as they are stationary. The challenge here is how to make a variable stationary at a value that is less than 1. That is, instead of having a variable stationary at I(1) we should as well be able to have the variable stationary at I(<1) in order for the variable to retain some memory vital during model development. To achieve this, we propose a technique that detects the optimal value of K - the least value of K - that makes a variable stationary without causing much memory loss.

In addition to developing an approach that helps to detect the optimal value for making a variable stationary, we also develop predictive models based on two datasets. The first dataset is generated from variables that have been made stationary using the popular approach of log differencing. The second dataset has been processed using our fractional differencing approach. We compare the results in order to examine if it is worthwhile to put forward our proposed approach of fractional differencing based on the model results. The score metrics considered are F1-Score, precisions, recalls and accuracies. All these metrics are higher from the model developed from our proposed approach when compared with the model results from the dataset processed using log differencing. This is a very notable contribution to our work.

Interestingly, the models we have developed have also performed well in predicting the directions of the stock market with an accuracy of 70%. We follow the same approach by [105] in identifying events based on time-varying volatility event-sampling. We introduce some popular technical analysis indicators into the model and our findings show that these indicators do improve our developed models as shown in Figure 6.5.1.

It would be interesting to incorporate sentiment elements into this model in

order to see if it can help to detect events more effectively. We understand that this model is applicable in Options-like trading in that it considers just two time points e.g. the price at the time one enters trading and the price at the time one exits the trading position. Only these two prices are considered, and the directions are predicted based on these two time points. This approach can therefore be extended to also cover spread betting and contracts for difference (CFDs) whereby all prices that occur between the entry and exit points are taken into account - the cumulative sum of returns.

7

Event-based algorithmic intraday trading with applied natural language processing algorithms

7.1 Motivation

So far, in previous chapters, we have explored daily sentiment data that have already been processed by third parties to examine the statistical influence of sentiments on the stock market.

In this chapter we aim to establish the causal relationship between sentiments and the stock market without reliance on any processed sentiment data from a third party. To do this, we employ our proposed and trained BERT-based NLP model introduced in chapter 5 to extract sentiment polarities from the financial

news collected.

We develop a high frequency stock market predictive model by combining the traditional stock market predictors used in chapter 6 and the extracted sentiment polarities from the financial news collected to see if the model with sentiment variables would outperform the model without the variables.

7.2 Introduction

The stock market is considered to be connected with many other sectors of the economy. In chapter 2 for example, a reference was made where former US president Donald Trump announced a tariff increase on steel imports in March 2018. This announcement caused disruptions in steel-related stock indices as a result of the sentimental contagion in the stock market. It led to the downturn in steel-related stock prices. Clearly, this had nothing to do with the fundamental stability of the industry. This view was also shared by Zhong and Enke [140] by stating that the stock market is affected by many interrelated factors such as economic, social, psychological and company-related variables. All these factors contribute to the highly ups-and-downs market volatility ([116],and [79]).

The causal relationship between sentiments and the stock market has been heavily studied in many researches and also covered in chapters 2, 3 and 4 with findings validating the causal relationships between them.

The sentiment variables captured in the previous chapters are processed daily time series and so are the stock market predictors - all the variables are time-sampled at a constant daily time interval.

In this work we explore intraday data. Doing this will help us to understand the time sensitivity of the sentimental information on the stock market. As opposed to the previous chapters, we attempt to extract the sentiments directly from the financial news by employing our proposed optimised BERT-based NLP algorithm developed in chapter 5. The financial news data collected relates to the constituents of the S&P 500 stock index. Presented in Figure 7.2.1 are the trends

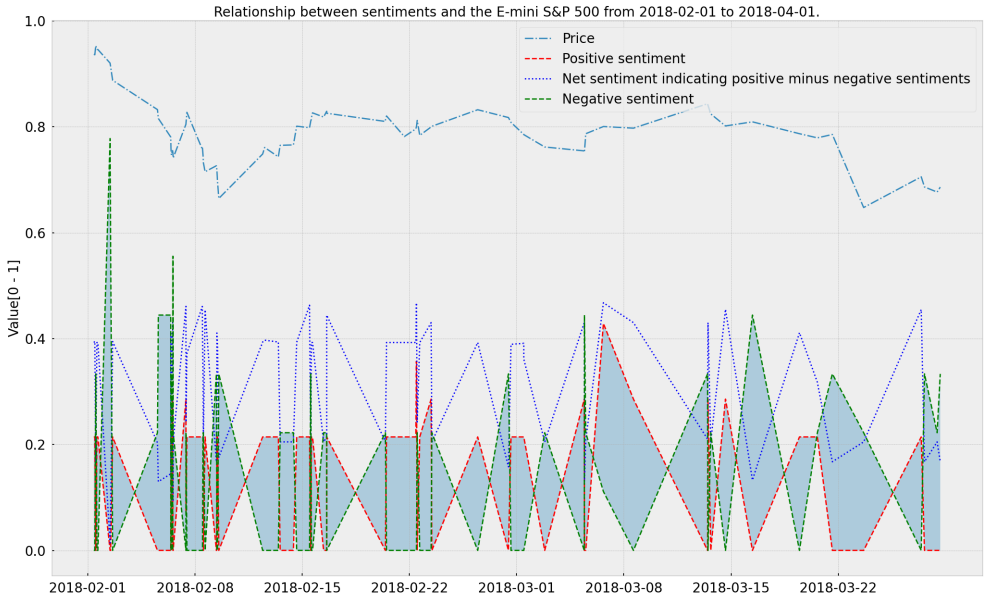

Figure 7.2.1: Graphical representation of the relationship between sentiments and the S&P 500 E-mini prices.

from the stock market and the sentiment indicators.

All the predictors have been normalised to values between 0 and 1. The topmost trend in the Figure indicates the stock price trends. As presented in Figure 7.2.1 the relationship between the sentiments and the stock market is hard to decipher by mere looking at their trends. On the one hand, looking at the net sentiment trend, we could see that the occurrences of the troughs and peaks coincide more frequently with the price trend's. This, of course, does not provide any credible substantiation to conclude that sentiments do influence the stock market. On the other hand, fluctuations in the price trend seem to have occurred much more frequently than they did in the sentiment trends. This observation could as well invalidate the possible relationship between the sentiments and the stock market. The frequent fluctuations would be normal expectations in the stock market because of its highly volatile nature. Establishing the relationship between the sentiments and the stock market is of interest in this chapter especially in view of high frequency trading where the stock price data is reported in nanoseconds -

we would not expect the sentiment data to occur in nanoseconds.

Clarity would be provided in the next section about how the sentiment variables are derived. Extracting the causal relationship between these variables is therefore the centrepiece of this chapter.

We aim to establish the causal relationship between the sentiments and the stock market and also to predict the directions of the stock market returns. In the process we propose a predictive model based on the integration of the stock market indicators, technical analysis indicators and the sentiments extracted from the financial news that are related to the constituents of the S&P 500 stock index. Indeed, this study is the first work to have employed a BERT-based NLP model for extracting sentiments from the financial news and incorporated the derived sentiment polarities, technical analysis indicators and stock market variables in building a stock market model for predicting the directions of the stock market. The contributions of this chapter are highlighted below:

1. We employ our proposed fractional differencing technique to the intraday stock market and technical analysis variables in order to achieve variable stationarity property.

2. We derive the sentiment polarities from over 1 million financial news related to the constituents of the S&P 500 stock index by employing our proposed BERT-based NLP model developed in chapter 5.

3. We develop an event-driven machine learning predictive model that depends on the stock market volatility, intraday stock market data, technical analysis indicators and extracted sentiments in predicting the directions of the stock market.

The novelties of the work in this chapter are centred on how sentiments are extracted from the financial news related to the constituents of the S&P 500 stock index and how the non-stationary variables are made stationary by employing our developed fractional differencing technique that optimally differences variables in order to achieve stationarity.

The remainder of this chapter is organized as follows. Section 7.3 presents our data preprocessing methodology. Section 7.4 details the machine learning model introduced and the results obtained. Our findings about the causal relationship between sentiments and the stock market are provided in this section. Finally, Section 7.5 discusses our findings and concludes this chapter.

7.3 Stock data and sentiment information

Three datasets are explored in the process of developing our stock market predictive model. Two of the datasets are obtained from the S&P 500 E-mini stock index. They are the stock market variables and the technical analysis indicators. The last dataset would be sentiment variables extracted from the financial news. The financial news data is sourced from a news aggregator platform [60] and is related to the constituents of the S &P 500. The data collected covers the period between September 2012 and July 2019 and we have 1.05 million records. Figure 7.3.1 illustrates how the datasets employed are generated from the two main sources.

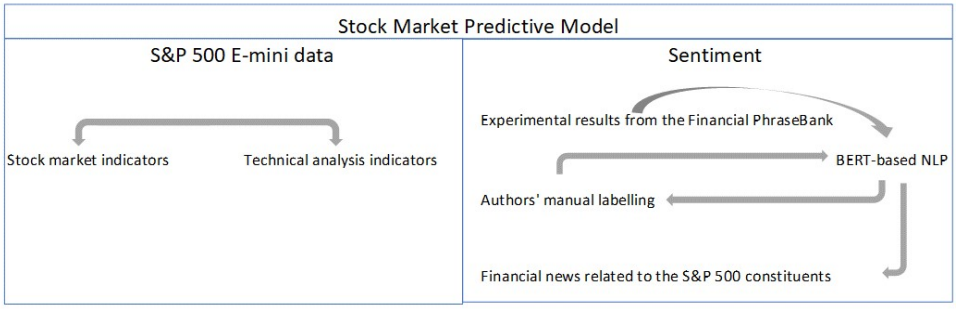

Figure 7.3.1: This information captures the datasets explored. The stock market variables and the technical analysis indicators are both extracted from the E-mini S &P 500 stock data. The sentiment dataset is derived from the financial news related to the constituents of the S &P 500. The experimental results from the Financial PhraseBank are obtained from [62] with more detail in chapter 5

To simply put, the focuses of this chapter would be to examine the statistical relevance of the sentiment information on the stock market and predict the stock market directions using our proposed predictive model. The predictive model developed in chapter 6 employs the stock market variables and the technical analysis indicators to predict the directions of the stock market. In order to evaluate the impacts of the sentiments on the stock market we develop a model that includes the same data employed in chapter 6 and, in addition, we include sentiment variables. This can be generalised as:

$$Accuracy_1, Recall_1, Precision_1 : Model_1 = \alpha_1 + \beta_1 Stock + \gamma_1 TechIndicator + \varepsilon_1 \quad (7.1)$$

$$Accuracy_2, Recall_2, Precision_2 : Model_2 = \alpha_2 + \beta_2 Stock + \gamma_2 TechIndicator + \eta_2 Sent + \varepsilon_2 \quad (7.2)$$

where *Stock, TechIndicator* and *Sent* denote the stock market variables, technical analysis indicators and sentiment variables respectively. The models $Model_1$ and $Model_2$ would be used to measure the influence of the sentiments on the stock prices. The difference in the models is that $Model_1$ does not include the sentiment variables - it only uses the stock market variables and technical analysis indicators. $Model_2$ adds the sentiment variables to the $Model_1$'s variables. If $Model_2$ performs better than $Model_1$ judging by the values of their accuracies, recalls, and precisions, one could conclude that the sentiments have predictive information on the stock market. $Model_1$ represents the model employed to generate the findings presented in Table 6.5.5 because the model was developed based on the stock market variables and the technical analysis indicators. We would therefore be developing a model that is similar to $Model_2$ and the findings from this model would be compared with the findings from $Model_1$ captured in Table 6.5.5 with respect to their values of accuracies, recalls and precisions. The

details of the methodology behind the stock market variables and the technical indicator indicators are provided under the methodology section 6.3 of chapter 6. Regarding how the sentiment variables are derived we present the details in the subsection that follows.

7.3.1 Sentiment variables

Recalling the Efficiency Market Hypothesis (EMH) that states that the market is efficient because stock market prices already reflect all known information and as a result, it would be hard to profit from any active trading engagement ([37]).The implied interpretation of this concept is that regardless of the amount of effort invested in the stock market trading active and passive traders earn almost the same returns. If one should place a value/cost on the level of effort invested by these traders, it would mean that the active traders would be worse off because of their cost of effort. But one thing is unclear regarding this hypothesis: how much time does it take for the arrival of new information to reflect in the stock market? Could it be microseconds, seconds, minutes, hours, or even days?

Grossman and Stiglitz [47] looked into the EMH by constructing a simple model of a futures market that was considered to be informationally efficient - the model assumed there was a large number of farmers, each of whom had perfect information about his own crop, but with little or no information about the crops of others. But the role played by the equilibrium price on the futures market coupled with the assumption that all farmers had the same constant absolute risk aversion utility function would help to aggregate this diverse information and make the market efficient.

But the assumption of this constant risk aversion utility mentioned in [47] seems impractical. Traders in general have disparate levels of risk aversion. As a result, the market is informationally inefficient. This is later supported by Gale and Stiglitz [65] by concluding that markets cannot be informationally efficient, in the sense that prices convey all of the information of the informed to the uninformed.

Many researchers disagree with the EMH, but there seems to be a very thin line where they all seem to have one thing in common with the hypothesis - market reactions to the arrival of new information. How fast does the stock market react to the arrival of new information? Who are the informed and who are the uninformed? How long does it take the informed to inform the uninformed?

To answer these questions, we investigate the sentiment analysis of financial news to see if the sentiments extracted from the financial news gathered would be statistically significant in the stock market modelling and, if so, how relevant is the time sensitivity of the sentiments? Answers to these questions would help to clarify on the impacts of sentiments on the stock market as a whole and uncover if there are first-mover advantages in the first place.

In the process of conducting sentiment analysis on financial news we source for financial news data from [60]. The data collected covers the period between September 2012 and July 2019 and it comprises over 1.05 million records. The financial news data is associated with the constituents of the S &P 500 index. The reason for limiting the data collection scope to cover just the constituents is because of the stock market data of interest which is the E-mini S &P 500 futures used in our stock market predictive model development.

We use our trained BERT-based NLP model developed in chapter 5 to extract the sentiments from the financial news. This is achieved by following a two-step approach of training the optimised BERT model introduced by Liu et al. [135] with the experimental results on the Financial PhraseBank data containing the 2263 financial news with 100% agreement level labelled by 16 financial experts. The second step is to use the trained BERT model to validate the manual labels of the 3000 financial news sampled from [60] (please see chapter 5 for details of our trained BERT-based NLP model). This model is then used to extract the sentiments from the financial news we have gathered from [60]. Four sentiment variables are generated in this process.

1. Positive sentiment is the count of all the financial news classified as positive in constant time interval of 1 second.

2. Negative sentiment is the count of all the financial news classified as negative in constant time interval of 1 second.

3. Neutral sentiment is the count of all the financial news classified as neutral.

4. Net sentiment is the 10-period exponential moving average of the difference between the counts of the positive and negative sentiments (Positive sentiment - negative sentiment).

In the final model we only include the net sentiment denoted as *Sent* and its lags $Sent_t$ where $t = 1, \ldots, 5$. Table 7.3.1 contains the list of the sentiment variables employed in the model.

Table 7.3.1: List of the sentiment variables explored.

Index	Variable name
1	*Sent* indicating the net sentiment at the current time
2	Lags of *Sent* ($Sent_1, Sent_2, \ldots, Sent_5$)

7.3.2 Stationarity

Some of the variables in our dataset are non-stationary. Instead of applying the traditional technique of log differencing we follow the proposed approach we introduced in chapter 6 by fractionally differencing them. Findings from chapter 6 have shown that the fractional differencing of variables is more effective for achieving stationarity as it helps the variables to retain their memory and consequently improve on their predictive power.

Our approach has been developed in such a way that it would optimally identify the best value needed in making the non-stationary variables stationary. For example, when a non-stationary variable is log differenced and becomes stationary, we say the variable is integrated at order 1. We could denote this as

$I(K = 1)$. But our proposed approach examines if a variable is stationary. If found not to be, it identifies the optimal value K that makes the variable stationary at $K < 1$.

7.4 Model development and applications

We would like to compare the results from the two models $Model_1$ and $Model_2$ in Eq. (7.1) and Eq. (7.2) respectively. We have the results for the model $Model_1$ as presented in Table 6.5.5. Therefore, we would need to establish the results for the model $Model_2$. This would imply that we combine the three datasets in Tables 6.3.2, 6.3.3 and 7.3.1 as defined in the general expression of the model $Model_2$.

Having employed all the relevant datasets to a simple SVM model the findings are therefore presented in Table 7.4.1.

Table 7.4.1: Support vector classification evaluation results from the fractionally differenced (I(<1)) test and validation datasets.

	Balanced training dataset			
Label	Precision	Recall	F1 Score	Support
-1	0.57	0.64	0.60	161
1	0.74	0.69	0.71	246
Test sample accuracy		0.67		
Validation accuracy		69		

Interestingly, the results show that there is no statistical significance between $Model_1$ and $Model_2$ on account that the values of their recalls, precisions, and accuracies are relatively the same. The implication is that the sentiment variables incorporated into the model $Model_2$ do not add any statistical improvement to the model. In other words, they are irrelevant in the model development. For a clearer understanding of their relative relevance with respect to other variables Figure 7.4.1 presents the feature importance of the final list of the variables used in the model. The sentiment variables appear to have the least importance

relatively.

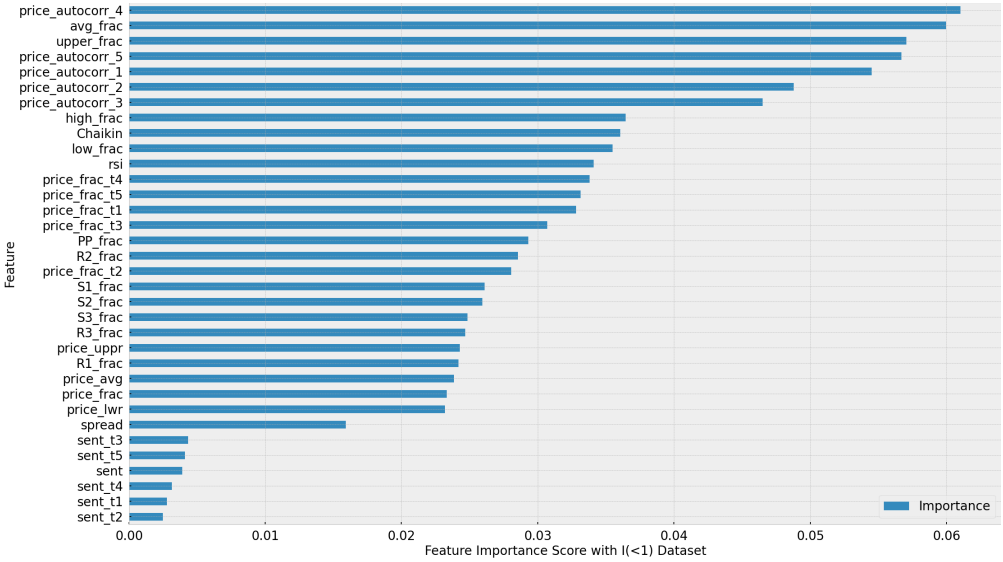

Figure 7.4.1: Feature importance of the variables explored in the predictive model development.

The features importance of the variables is presented in 7.4.1 where the sentiment variables appear to be the least important in the model. Are we to generally conclude that sentiments do not have any statistical significance with respect to the stock market based on the findings from this chapter? There are many factors that we need to consider first. Among the factors are the following:

1. Is the financial news gathered complete?

2. Is there any time delay before the news is reported?

First, many stock market automation algorithm developers source financial news data from multiple and popular news aggregators like Bloomberg and Reuters. These news aggregators have been around for a long time and with years of experience. So, it is highly questionable if the financial news explored in our model is complete because the news was sourced from a single aggregator.

Another important factor to consider is the potential delay in the financial news gathered. Are the financial news explored in our chapter the first source or just a repetition of news by other financial news sites? If the news sourced are not aired for the very first time, then they might have lost their economic influence on the stock market. In other words, the stock price might have already reflected the information from the financial news.

It was observed in Figure 7.2.1 that fluctuations occurred much more in the stock price trend than from the sentiment trends. This is very understandable because of the volatile nature of the market. And the stock market data is re-sampled using our introduced event-driven CUSUM technique that depends on the stock market volatility. These events may not necessarily match well with the sentiments generated in the stock market. For example, support and resistance indicators from the technical analysis react strongly to the forces of demand and supply in the stock market and they are there to control the stock market prices.

In view of these factors, it appears hard to conclude that sentiments do not have statistical influence on the stock market. Sourcing financial news from multiple sources, eliminating duplicates from financial news, and sampling the stock market data-based events from the financial news may help to reveal the true relationship between sentiments and the stock market.

7.5 Discussion and conclusion

Findings from many research works have attempted to examine the relationships between sentiments and the stock market. Establishing the true relationship requires complete care and attention to detail with respect to many factors. The stock market is full of many unsurprising surprises and we would expect to come across many frequent ups and downs. By nature, the stock market is highly sensitive not just to the fundamental values of their underlying stocks but also to many other factors that seem to relate to the stock market. One of the most noticeable examples is the political factor. This is just an example of how highly

sensitive the stock market is to a myriad of factors. As a result, we would expect maximum care in the process of developing a stock market predictive model for predicting the directions of the stock market.

We made an attempt to examine the relationship between sentiments extracted from financial news based on our developed optimised BERT-based NLP model and the event-driven stock market sampling based on the CUSUM technique and technical analysis indicators. Findings from our model reveal no significantly statistical relationship between them. However, we express some limitations about the source of the sentiments generated and how they were exploited in the model. These concerns may have biased the findings from our model.

For future work, we suggest a sentiment-driven stock market intraday data sampling whereby data sampling occurs more frequently during high sentiments as opposed to event-driven sampling that relies on high return volatility. We would also like to consider sourcing sentiment data from multiple and reliably popular sources such as Bloomberg and Reuters for example. This may help to reveal the true relationship between sentiments and the stock market.

8

Contracts for Difference (CFDs) machine learning algorithm for optimising portfolios

8.1 Motivation

Machine learning is a fast-growing trend in many industries. Its applicability covers areas such as corporate finance, retail, banking, and health, among others. More specifically, stock market traders are leveraging machine learning algorithms in making informed decisions that include investment and strategic trading. Trained automation algorithms automatically detect opportunities and act on them much faster than humans can. Interestingly, the literature related to these predictive algorithms on contracts for difference (CFDs) is almost

non-existent. Clearly, the stock market is a very sensitive environment known for its extremely volatile nature. This chapter ventures into developing CFD-focused predictive algorithms and evaluates their efficiency in portfolio optimisations.

8.2 Introduction

High frequency trading (HFT) has become popular in the capital markets around the world. HFT has been brought to life with the advancement in big data and high-speed computing technology. It involves developing a computing algorithm embedded with instruction parameters to carry out automated trading strategies at high frequency with no human intervention. HFT traders are known as algorithm traders that trade via electronic systems. They rely heavily on high-speed computing technology to connect them to trading platforms for placing and executing orders. Among the popular products traded by these algorithms traders are financial security derivatives such as ETF futures and contracts for difference. These are mainly arrangements between parties where the differences in the settlements between the open and closing trade prices are cash-settled. For simplicity, let us assume that a trader holds a long position with a futures contract value of $\$\ 5000$ at time t_1. The investment time horizon is denoted as t_1, t_2 and t_3 with stock prices at $\$40$, $\$43.50$ and $\$42.50$ respectively where $t_3 > t_2 > t_1$. At the exit time t_3, the trader would have made $\$125$ generated from $((43.50 - 40) + (42.50 - 43.5)) * 5000/100$.

The question that comes to mind is how to develop HFT algorithms on futures trading that can guarantee us positive returns. Despite the extensive success of machine learning models in strategic decisioning, their literature in CFD securities seems non-existent. One would expect any machine learning algorithms for this aspect of the capital market to help maximise profits while minimising investment risks. The asks and expectations from these algorithms are very high, ranging from their statistical appropriateness, perceptiveness, and insightfulness in portfolio optimisation. But there are frequent ups and downs and expected surprises in the stock market. And the stock market is highly

susceptible to a myriad of events. Understandably, in the face of these challenges developing a reliable machine learning algorithm that can satisfy these high expectations would be hard to come by; hence, the lack of literature presumably.

In this chapter we take a bold step to lead the literature in the applications of machine learning models in the HFT futures market. To start with, we introduce a novel event-driven forward labelling approach for extracting insights from the historical HFT ETF data and training a machine learning model with the insights learnt with the aim of optimising our investment portfolios. Our proposed model would be able to automatically detect opportunities in the HFT-EFT market, advise on the trading positions to take, and suggest when to exit the positions. This is similar to the goal of the meta-labelling technique introduced by [105] in that our proposed approach helps to identify a trading position for a given data point based on the patterns learnt. This would be the first work that has boldly endeavoured in this research direction. Our proposed event-driven forward labelling technique incorporates the CUSUM with the expected cumulative sum of returns in identifying the informative patterns during the model development. The event-driven sampling is required for identifying the events in the stock market, but it does not constitute making strategic investment actions in isolation, but when combined with the expected cumulative sum of n returns, it helps in the labelling of the training dataset during the model development.

We explore an OXGboost framework for developing our stock market predictive model. And the focus would be that we predict the directions of futures contracts which are also similar to contracts for difference. The contributions of this chapter are highlighted below:

1. We introduce a novel event-driven forward labelling technique for identifying the patterns in the stock market trends and labelling the training data for model development.

2. We develop a machine learning predictive model for predicting the directions of the CFDs. The literature on applying machine learning models to CFDs is almost non-existent. This is clearly a bold step we have

taken in this direction.

3. We propose a novel HFT predictive model that incorporates both the binary and multi-class OXGboost frameworks. Our proposed model is designed to detect trading opportunities and action on them accordingly. This model would advise regarding when to enter a trading position, what position to take and when to exit the position with the purpose of optimising investment portfolios.

The remainder of this chapter is organized as follows. Section 8.3 presents the data explored and introduces the event-driven forward labelling proposed in this chapter. The empirical findings are presented in section 8.4. Finally, section 8.5 ends this chapter with a summary and discussion.

8.3 Methodology

We use the same ETF data explored in chapter 6 except that we explore a longer data period. The data is sampled at a constant time interval of 1 second of time and it covers the period between 5th of October 2009 and 12th of June 2020. The details of the data, the variables obtained, and how they have been pre-processed using our novel stationarity technique are provided in chapter 6.

For the independent variables we use all the variables established in chapter 6. The proposed model in chapter 6 focuses on Options market. But in this chapter we are considering futures ETF trading like CFDs. So, we would diverge away from the labelling approach employed in chapter 6. Instead, we introduce a novel labelling approach termed as the event-driven forward labelling approach. The details of this approach are presented in the subsection that follows.

Event-driven forward labelling

In futures trading one could be in any of the following position: long, short or hold positions. With the long position expectation is that the cumulative sum of every price change between the point of entry and the point of exit is positive. If

this is the case, then a futures trader would realise positive returns. This implies that on average there is a stock price increase, and this generates investment returns. On the other hand, we expect the cumulative sum to be negative where a short position is held if one is to realise profit when there is a price fall on the average. And for the hold position, no trading action is undertaken. The challenge here is to know how to label the data so that the three scenarios are captured. And this leads us to introduce our proposed event-driven forward labelling. This is an advancement to the CUSUM technique and a diversion to the meta-labelling approach introduced by [105] in view of the different financial instruments under consideration.

Given n as the forward looking number of futures ETF transactions, and corresponding to the number of transactions is the forward looking cumulative sum representing the return on investment denoted as r_n. We therefore present below the condition upon which the label indicating an investment position is assigned:

$$y_t = \begin{cases} -1, & \text{if } -r_n > -r_{2xn} > -r_{3xn} > -r_{4xn} \\ 1, & \text{if } r_n < r_{2xn} < r_{3xn} < r_{4xn} \\ 0, & \text{if } else \end{cases} \tag{8.1}$$

where $2xn$ implies multiplying the number of transactions n by 2. For example, for $n = 360$, $2xn$ would result in 720 forward looking transactions between the trading points of entry and exit. r_{720} would therefore be the forward looking expected return (cumulative return) after 720 transactions. Appendix A.0.8 is the Python function for our proposed event-driven forward labelling.

Table 8.3.1 demonstrates how transactions are labelled based on the Eq. (8.1). As we can see that only the transactions with forward looking expected returns that are monotonically increasing and decreasing are labelled as 1 and -1 respectively. Even when all the returns for the 4 sets of forward looking transactions - expected returns at 360, 720, 1080 and 1440 transactions - are just positive or negative, they are still labelled as 0. For transactions to be labelled as 1

we would expect that the conditions stated in Eq. (8.1) are satisfied. That is, all the four cumulative sums of returns are positive and monotonically increasing. The reverse is the case with label -1.

Table 8.3.1: An illustrative example of how the event-driven forward looking labelling assigns labels where $times = 4$ and $factor = 360$ as defined in Appendix A.0.8. "$>$" denotes the consecutive cumulative returns are monotonically increasing and "$<$" implies the reverse.

Transactions	360	720	1080	1440	Description	Label
Expected return1	0.54	2.26	1.67	7.04	Positive	0
Expected return2	-0.54	-2.26	-1.67	-7.04	Negative	0
Expected return3	-0.54	2.26	1.67	-7.04	Unstable	0
Expected return4	2.96	4.39	6.4	7.04	$>$	1
Expected return5	-2.96	-4.39	-6.4	-7.04	$<$	-1

Model methodology

Our interest is to be able to identify and exploit trading opportunities. That is, we would like to know when we can trade, what trading positions to take, and when to exit them. So, the labels 1 and -1 indicating long and short respectively represent the main points of interest. We would expect label 0 which indicates no trading action to dominate the whole transactions. As we are seeking opportunities in the stock market and with the thousands of transactions per day in the HFT-ETF market, the trading opportunities we are interested in occur very rarely - we would not be surprised if they occur once or twice in hours or days. We would also expect a likelihood of not identifying any trading positions for days as well.

The parameters in Eq. (8.1) are pre-defined as follows:*times* is assigned the value of 2, the sampling frequency *freq* is set as hourly and the initial cumulative sum factor is assigned the value of 100. The frequency of sampling being set as

hourly is to ensure uniqueness in the sample data gathered for the model development - we do not want to have any overlap in the training data among some data points.

In order to avoid some information leakage between the training and out-of-sample datasets we stay away from shuffling the data during the data splitting. Training dataset is in the period from the 28th of September 2009 to the 12th of June 2019. The out-of-sample dataset is for the period from the 13th of June 2019 to June 2020.

Table 8.3.2: Distribution of trading position labels.

Label	*Count*	Percent
0	6906	74%
1	1337	14%
-1	1087	12%

Having applied our proposed event-driven forward looking labelling approach in Appendix A.0.8 on the EFT data the distribution of the labels is presented in Table 8.3.2. Most of the data is labelled as value 0. Now, the question is how to identify opportunities. This leads us to our proposed predictive model that incorporates both the binary and multi-class OXGboost frameworks.

First, we develop a binary classification algorithm that makes use of only the labels 1 and -1 obtained by applying our proposed event-driven forward labelling technique. That would imply that the label value 0 is filtered out having applied our proposed labelling technique. In this process we ensure that only the data with extreme labels 1 and -1 is used in training the binary model. Second, we develop a multi-class model that makes use of the 3 labels generated. This is illustrated in Fig. 8.3.1.

With the multi-class model, we identify the long and short trading positions. In addition, we use the outputs predicted as long and short positions from the

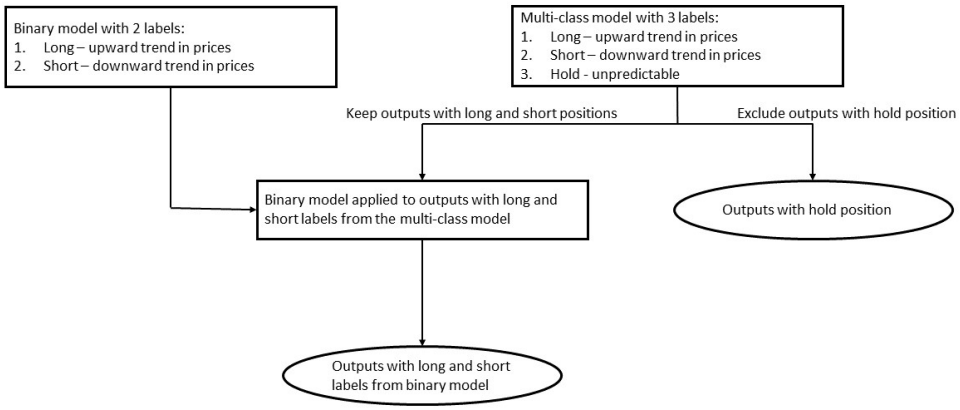

Figure 8.3.1: Proposed OXGboost Model Architecture.

multi-class model as inputs in the step-2 binary model to predict the directions of the stock market. The benefit of adding the binary model is to complement and improve on the results from the multi-class model.

In the next section our proposed OXGboost model is applied to the ETF data and the findings are presented.

8.4 Model application and findings

In the process of developing our proposed predictive model, the ETF data is split into 2 datasets: the training and out-of-sample datasets.

For the binary model, the label value 0 is excluded in the model development. We employ OXGboost to the training dataset with stratified 10-fold cross validation and some parameter fine-tuning. We evaluate the model performance on the out-of-sample dataset at some future times. The results obtained are presented in Table 8.4.1

As observed from Table 8.4.1 the levels of accuracy obtained from the training dataset is similar to that of the out-of-sample dataset's. The out-of-sample dataset is ahead in time of the training dataset. This shows that the model is reliable because of the consistency in the model performance between the two datasets.

Table 8.4.1: Binary classification model. We employ stratified 10-fold cross validation. The best parameters identified are the following: subsample 1.0, min child weight 1, max depth 4, gamma 0.5, and colsample bytree 0.8. Training dataset is between the 28th of September 2009 and 12th of June 2019. The out-of-sample dataset is between the 13th of June 2019 and June 2020.

	Binary OXGboost model results			
Label	Precision	Recall	F1-Score	Support
-1	0.58	0.66	0.62	157
1	0.71	0.63	0.66	203
Out-of-sample accuracy		0.64		360
Training accuracy		0.64		2024

Our interest would be that we correctly identify strategic trading positions such as the long or short positions. We would also expect that the hold positions indicating trading inactivities to be disproportionately high in relation to the other labels. We may also be faced with the situations that no strategic trading positions are identified in days. In order to identify and filter out the trading inactivities we also develop a multi-class model and the results are presented in Table 8.4.2

Table 8.4.2: Multi-class model. We employ stratified 10-fold cross validation. The best parameters identified are the following: gamma 0, learning rate 0.01, max depth 16, n estimators 400, reg lambda 1, subsample 0.7.

	Multi-class OXGboost model results			
Label	Precision	Recall	F1-Score	Support
-1	0.21	0.56	0.30	157
0	0.67	0.31	0.43	685
1	0.27	0.39	0.32	203
Out-of-sample accuracy		0.36		1045
Training accuracy		0.55		3261

It can be seen from Table 8.4.2 that the multi-class model performs very poorly

judging by the levels of accuracy, precisions and recalls. Many features with actual label 0 are incorrectly predicted as 1 and/or -1. For a clearer view this is also presented in Table 8.4.3.

Table 8.4.3: Confusion matrix results from the multi-class model on the out-of-sample dataset.

	Predicted		
Actual	-1	0	1
-1	80	49	20
0	270	214	201
1	65	58	80

Before we introduce the trained binary model to correct the mismatches between the actual labels and the predicted labels, we relax the strict labelling conditions. Below are the strict conditions:

1. Consecutive cumulative sums to be monotonically increasing for label 1, decreasing for label -1 and label 0 otherwise.

2. The initial cumulative sum of return must be at least 0.2 for label 1, -0.2 for label -1 and label 0 otherwise.

Each of these conditions is necessary but not sufficient in the labelling process. That is, both conditions must be satisfied. So, we introduce a new condition to the labelling of the out-of-sample data before applying our trained binary model to the output of the multi-class model having filtered out outputs with label 0. This condition is applied so that the old conditions are relaxed and the new one is used to label the out-of-sample data. Below is the new condition applied:

1. Cumulative sum at the exit point must be $>$ 0 for label 1 and $<$ 0 for label -1.

Let us assume that we are interested in the cumulative sum of the 100 transactions, and that labels are assigned based on if the cumulative sum of return

is positive and negative, and that we only select the informative features predicted as 1 and -1 from the multi-class model. Then, the new condition is applied to the data before using the outputs from the multi-class model as inputs in the binary model. This process flow is employed to the binary model and the out-of-sample data and our findings are presented in Table 8.4.4.

Table 8.4.4: Trained binary model incorporated. The model uses the outputs with the predicted labels 1 and -1 from the multi-class model as inputs.

	Results from our proposed model.			
Label	Precision	Recall	F1-Score	Support
-1	0.56	0.72	0.63	329
1	0.69	0.53	0.60	390
Out-of-sample accuracy		0.62		719

As it appears in Table 8.4.4 we have a good model that is able to predict the directions of the CFDs securities. This has been achieved, in part, by relaxing the conditions initially applied during labelling. Also, the trained binary model is applied to the outputs of the multi-class model.

8.5 Discussion and conclusion

This chapter is centred on developing a stock market predictive model for CFDs securities which have become popular in the capital market. High frequency intraday data of 1 second is employed during model development.

We introduce a novel technique termed the event-driven forward labelling that helps to sample the available high frequency intraday data during some stock market events, filter out the data points with no informative features, and support with the labelling of the relevant data points. This technique incorporates the popular event-driven CUSUM method, puts forward a list of conditions that includes expected consecutive cumulative sums of returns with monotonic trends and minimum expected returns.

We employ the OXGboost framework that combines the binary and multi-class models. The main aim is for the model to help in detecting trading opportunities, advising on what strategic investment positions to take for achieving profit maximisation based on longing and shorting.

The findings from our proposed model are promising. First, the model is able to identify strategic investment opportunities in the CFDs trading market, advise on the trading positions to take, and suggest when to exit the positions based on the expected cumulative sums of returns. This would imply that by employing the model, traders can benefit both from shorting and longing in the capital market. It is well known that the chance of profiting in the stock market is extremely low. According to the Financial Conduct Authority (FCA) [59] over 82% of people involved in betting and CFDs trading lose money based on a sample of industry data, and as a result, the FCA has raised significant conduct concerns about the amount of leverage being offered.

So, having a model with a high level of accuracy is promising. Second, the trained model is applied to the data that the model has not seen before, and the model still performs very well in terms of its accuracy, recall and precision.

As a future work, it would be interesting to see how the proposed model will work in real a life situation.

9

Conclusion

With knowledge of how highly volatile the stock market is this thesis examines the statistical significance of sentiment on the stock prices. More specifically, it studies the causal relationship between sentiment and the stock market with the goal of developing a stock market model for predicting the directions of the stock market prices.

So far, the literature within this purview has offered conflicting conclusions about how sentiment influences the stock market. The scope of this thesis therefore covers the clarifications of the relationship. First, it is noteworthy to understand that sentiment is a broad subject - it can be extracted from many sources such as social blogs to represent public mood, financial news to indicate financial expert opinions, among others. Different sources of sentiment information are explored in order to understand the relevance of the sources of sentiment with respect to the stock market.

In addition, we examine the time sensitivity of sentimental information by exploring daily and very high-frequency sentiment information and establishing their relationships with the directions of the stock prices.

Equally important is the applicability of the right model selection. We show that model selection plays a key role in validly establishing the relationship between sentiment and the stock market. The selection of a wrong model could bias model results and make them completely misleading. This thesis proposes some novel approaches that cover areas such as variable stationarity, data dimensionality reductions, model robustness, and market direction predictions.

We explore the stock market data from various sources and structures. For example, daily-based and event-driven intraday models that predict the directions of the stock market prices are developed. Included in this thesis is a novel optimised BERT-based NLP algorithm that extracts sentiments directly from financial news. The multi-class NLP model proposed shows very promising results judging by its high level of accuracy in identifying the sentiment polarities from financial news.

In the next section we provide the summary of the thesis in each chapter for a better understanding of the scope of the thesis.

9.1 Summary of the thesis

One of the focuses of this thesis is the examination of the causal relationship between sentiments and the stock market. This area of interest is covered in most of the chapters but with different data sources, methodologies, model selections, among others. We also develop stock market predictive models that predict the directions of the stock market prices. Both daily and high-frequency intraday datasets are explored. In this section we summarise the work presented in each chapter and emphasize on their respective contributions.

Chapter 2. Most of the previous studies claiming that emotions have predictive influence on the stock market do so by developing various machine learning predictive models, but do not validate their claims rigorously by analysing the

statistical significance of their findings. In turn, the few works that attempt to statistically validate such claims suffer from important limitations of their statistical approaches (Gilbert and Karahalios [38]). Stock market data exhibit erratic volatility, and this time-varying volatility makes any possible relationship between these variables non-linear, which tends to statistically invalidate linear based approaches. We therefore re-visited the work of [38] that relied on a linear framework which turned out to be unsuitable for accessing the causal relationship between the sentiments obtained from LiveJournal posts. Our work tackles these kinds of limitations, and extends linear frameworks by proposing a new, non-linear statistical approach that accounts for non-linearity and heteroscedasticity. Although our work has established that the Anxiety Index does not have predictive information with respect to the stock market, we observe some concerns as to how the Anxiety Index was built, based on incomplete data, non-specific LiveJournal posts, corpus challenges, non-representative data sample, among others. Further refining the process of defining the Anxiety Index by addressing the above-mentioned concerns, may help to fine-tune our empirical results and provide us with a more reliable predictive model.

Chapter 3. This chapter assesses the asymmetric impacts of positive and negative sentiments on the stock market returns by using a non-parametric nonlinear approach that corrects specific limitations encountered in previous related work. In addition, it proposes a new approach to developing stock market volatility predictive models by incorporating a hybrid GARCH and artificial neural network framework, and proves the advantage of this framework over a GARCH only based framework. Daily aggregated sentiments from StockTwits which contains sentiment-filled S&P 500 blogs on Twitter are used as the sentiment variables. Our results reveal that past volatility and positive sentiments appear to have very strong predictive power over future volatility. In conclusion, we emphasize on the importance of the source and suggest that one must pay attention to the source of sentiments used in developing stock market predictive models.

Chapter 4. Collective intelligence, represented as the sentiment extracted from social media mining, is encountered in various applications. Numerous studies involving machine learning modelling have demonstrated that such sentiment information may or may not have predictive power on the stock market trend, depending on the application and the data used. This chapter proposes, for the first time, an approach to predicting S&P 500 based on the closing stock prices and sentiment data of the S&P 500 constituents. One of the significant complexities of our framework is due to the high dimensionality of the dataset to analyse, which is based on its constituents and their respective sentiments, and their lagging. Another significant complexity is due to the fact that the relationship between the response and the explanatory variables is time-varying, and it is difficult to capture. We propose a predictive modelling approach based on a methodology specifically designed to effectively address the above challenges and to devise efficient predictive models based on Jordan and Elman recurrent neural networks. We further propose a hybrid trading model that incorporates technical analysis, and the application of machine learning and evolutionary optimisation techniques. We prove that our unprecedented and innovative constituent and sentiment-based approach is efficient in predicting S&P 500, and thus may be used to maximise investment portfolios regardless of whether the market is bullish or bearish.

Chapter 5. Some of the previous chapters that studied the causal relationship between sentiments and the stock market have explored daily pre-processed sentiment datasets obtained from third parties. As a result, we were able to avoid the challenges and complexities behind identifying and extracting sentiment polarities from financial news. But how reliable were the sentiment datasets explored? Until recently, NLP models such as the convolutional neural network and LSTM were common frameworks for sentiment analysis with models achieving their best accuracy of around 70%. Still, a good number of sentences are misclassified using these frameworks. Therefore, the reliance on such

processed sentiment data comes at its own cost. In fact, it calls into question the findings regarding the causal relationship between sentiment information and the stock market. Another concern is the use of daily aggregated sentiment data. Extracting sentiment polarities from financial news as they become publicly available could benefit from first-mover advantage. But these benefits may become eroded over time due to information symmetry. In view of these challenges, chapter 5 is centred on developing a BERT-based NLP model for extracting sentiments from financial news with the expectation of a much higher accuracy level. We proposed an optimised BERT-based NLP model and compared the results from our model with some notable works in NLP. Our proposed model shows very promising results and appears best in relation to the notable works revisited in the literature on NLP.

Chapter 6. We proposed a fractional differencing approach for processing non-stationary stock market variables in the process of developing a model for predicting the directions of the S&P 500 E-mini stock prices. On account of the high volume of intraday stock data, we introduced a CUSUM technique to help generate the event-driven data used during model development. The technique samples data more frequently during some events based on some time-varying return volatilities pre-estimated.

We examined and compared the effectiveness of the models developed with the two datasets processed using log differencing and our proposed approach of fractional differencing. The results show that the model trained with the variables stationalised using our proposed fractional differencing approach performs better because these variables retain some informative memory that helps them to improve their predictive power judging by the higher accuracy of 70% achieved.

Chapter 7. In chapter 7 three datasets were incorporated into developing a stock market predictive model. We used the traditional stock market dataset that has been generated using the CUSUM technique that helps to filter out features with no informative elements. We introduced some technical analysis indicators

as the second dataset. The third dataset covers the sentiment variables generated from financial news. The sentiment dataset was produced by using our proposed optimised BERT-based NLP model to extract sentiment polarities from over one million financial news related to the constituents of the S&P 500 stock index.

We attempted to examine the relationship between sentiments and the stock market. Presented in Figures 7.2.1 are some indications of commonalities in the trend movements between the sentiments and the stock market prices. But the findings from our model reveal no significantly statistical relationship between them. Upon further examinations it was observed that most values of the sentiment variables are zero. This makes the evaluation of their relationship impractical.

Chapter 8. Chapter 8 focuses on developing a model for predicting contracts for difference. Literature in this field seems non-existent. We take a bold step to lead in the literature by introducing a novel technique termed the event-driven forward labelling that helps to sample high frequency intraday data more frequently during some stock market events, filter out the data with no informative features, and support with the labelling of the relevant data points. Our proposed model would be able to automatically detect opportunities in the market, advise on the trading positions to take, and suggest when to exit the positions. The trained model is applied to the out-of-sample data the model has not seen before and it performs very well. This confirms the efficiency of our proposed model in the capital market where over 82% of traders often lose.

9.2 Summary of contributions

Overall, the thesis makes the following contributions:

1. It proposes a non-parametric statistical technique for detecting the nonlinear causal relationships between sentiments and the stock markets and this helps to avoid the limitations around the linear framework.

2. It introduces a new approach to developing stock market volatility

predictive models by incorporating a hybrid GARCH and artificial neural network framework and proves the advantage of this framework over a GARCH only based framework. In addition, it evaluates the asymmetric impacts of positive and negative sentiments on the stock market volatility.

3. It proposes, for the first time, an approach to predicting the S&P 500 based on the closing stock prices and the sentiment data of the S&P 500 constituents.

4. It introduces a hybrid trading model that incorporates a technical analysis, machine learning and evolutionary optimisation techniques for predicting the directions of the stock market prices.

5. It proposes an approach to reducing the number of dimensions, adapted to our framework, based on a 3-step approach, consisting of performing variable clustering, PCA, and by applying a variable selection method that we introduce here based on the modified Best GLM variable selection method initially developed by McLeod and Xu [6].

6. It proposes a novel optimised 2-step BERT-based NLP model for extracting sentiment polarities from financial news, and some findings from the multi-class model shows promising results judging by its high accuracy level in identifying sentiment polarities. A sample of the results from the developed model is presented in Appendix A.0.6.

7. It introduces a new method of achieving variable stationarity as opposed to log differencing for obtaining the Integration of Order I(1) in view of the fact that this long-preserved tradition diminishes the predictive power of variables. Instead, we propose a method that would first check for variable stationarity in a dataset, and then automatically detect and assign an appropriate optimal value to each of the variables in the dataset for them to satisfy the stationarity property. We build on the fundamental approach of the fractional differencing originally introduced by Hosking [72].

8. To the best of our knowledge, this thesis would be the first to extract sentiments from financial news based on our proposed optimised BERT-based model, and combine the results with technical analysis indicators and event-driven stock market variables sampled on time-varying return volatilities for predicting the directions of the stock market prices.

9. It proposes a new CFDs-focused predictive model that incorporates both the binary and multi-class OXGboost frameworks for automating trading strategies. This model is designed to advise on when to enter a trading position, what trading position to take and when to exit the position with the purpose of optimising portfolios.

9.3 Constraints and limitations

One of the main focuses of this thesis is the assessment of the causal relationship between sentiments and the stock market. To this end, different sentiment and stock market data sources have been considered and examined. As expected, each of the sources presents its own unique findings. As a result, we are unable to reach a very simple and categorical conclusion as to whether there is a Granger causality between the sentiment and the stock market variables. Regardless, we have some comforting resolutions leading us to conclude that data sources and model appropriateness play pivotal roles in assessing the Granger causality between them. However, there are some inherent limitations encountered ranging from the data sources, data validation, data availability, model development, among others.

We explored the relationship between sentiment and the stock market in Chapters 2, 3 and 4. Regarding the sentiment data sources, we have limited information with respect to the processing of the sentiments.

Another limitation is the restricted model selection and its level of accuracy covered. Chapters 2, 3 and 4 rely on processed sentiment datasets. Given that the

sentiments in the aforementioned chapters where processed before 2018 when a new powerful BERT-based NLP model had not been developed, we see this as another limitation of the sentiment datasets explored (Olaniyan et al. [111]).

More so, we observe some limitations around the limited model development approaches introduced. In Chapter 8, for example, we introduced a volatility-based event-driven technique to sample the stock price data used. Thereafter, we examined the Granger causality between the sentiment and stock market. This technique might have also weakened this relationship. It would be interesting to see what this relationship would be if we would introduce a sentiment-induced event-driven technique. Since this approach was not explored in our work, we consider this a limitation.

In Chapters 6, 7 and 8 we claim to have developed stock market predictive models. Applying these models developed in real life would have been the best way to evaluate their effectiveness, but we did not achieve this. This is seen as a limitation to this thesis.

9.4 Future research directions

Chapter 7 studied the statistical relevance of sentiments on the stock market. In the process, the event-driven stock market data is combined with the sentiments extracted from the financial news related to the constituents of the S&P 500 stock index and technical analysis indicators. Due to a significant proportion of missing values from the sentiment data, we were unable to identify the influence sentiment exerts on the stock market.

The causal relationship between sentiments and the stock market could have been better examined if the stock market data would have been sampled based on some events rather determined by the market sentiments as opposed to the stock market return volatility used. With this we could have easily compared the statistical relevance of the sentiment variables by looking at their variable importance and their contributions to the model predictions. This thesis can therefore be extended to address this issue.

References

[1] A. Baevski, S. Edunov, Y. Liu, L. Zettlemoyer, M. Auli: *Cloze-driven pretraining of self-attention networks*,arXiv preprint arXiv:1903.07785, 2019.

[2] A. Bechara: *The role of emotion in decision-making: Evidence from neurological patients with orbitofrontal damage*, Brain and Cognition,Development of Orbitofrontal Function, 55(1):30 – 40, 2004.

[3] A. Hatfield, J. Cacioppo, and R.L. Rapson: *Emotional contagion*, Cambridge University Press, 1994.

[4] A. Jahidul, A.H. Mohammad and H. Rajib: *Analyzing public emotion and predicting stock market using social media*, American Journal of Engineering Research, 2(9), 265–275, 2013.

[5] A. Marszalek and T. Burczynski: *Modeling and forecasting financial time series with ordered fuzzy candlesticks*, Information Sciences, 273, 144–155, 2014.

[6] A. McLeod and C. Xu: *bestglm: Best Subset GLM*, URL https://CRAN.R-project.org/ package=bestglm, 2010.

[7] A. Mittal and A. Goel: *Stock prediction using twitter sentiment analysis*, Project report, Stanford, 2012.

[8] A.M. Dai and Q.V. Le: *Semi-supervised sequence learning*, In Advances in neural information processing systems, pages 3079–3087, 2015.

[9] A. Parikh, O. Täckström, D. Das, and J. Uszkoreit: *A decomposable attention model*, In Empirical Methods in Natural Language Processing, 2016.

[10] A.T. Chen and M.T. Leung: *Regression neural network for error correction in foreign exchange forecasting and trading*, Computers Operations Research,31(7)1049–1068, 2004.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin: *Attention is all you need*, 2017, arXiv:1706.03762.

[12] A.Z. Zambom and R. Dias: *A review of kernel density estimation with application to Econometrics, arXiv:1212.2812v1*, 2012.

[13] B.J. Lobo: *Jump risk in the US stock market: Evidence using political information*, Review of Financial Economics, 8(2), 149–163, 1999.

[14] C. Alexander: *Market models.* 1st Edition, John Wiley & Sons, 2001.

[15] C. Blum,R. Chiong,M. Clerc,K. De Jong,Z. Michalewicz,F. Neri and T. Weise: *Evolutionary Optimization*, Variants of Evolutionary Algorithms for Real-World Applications, pp.1–29, 2012.

[16] C. Diks and V. Panchenko: *A new statistic and practical guidelines for nonparametric Granger causality testing*, Journal of Economic Dynamics and Control, *30(9–10)*, 1647–1669, 2006.

[17] C. Hiemstra and J.D. Jones: *Testing for linear and nonlinear Granger causality in the stock price–volume relation*, Journal of Finance, 49, 1639–1664, 1994.

[18] C.K. Reddy and C.C. Aggarwal: *Data Clustering*, Chapman and Hall/CRC,ISBN: 9781466558229, 2013.

[19] C. Sun, L. Huang and X. Qiu: *Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence*, arXiv:1903.09588, 2019.

[20] C.W.J. Granger: *Investigating Causal Relations by Econometric Models and Cross–spectral Methods*, Econometrica. 37 (3): 424–438, 1969.

[21] D. Araci: *FinBERT: Financial sentiment analysis with pre-trained language models*, arXiv:1908.10063, 2019.

[22] D.B. Nelson: *Conditional heteroscedasticity in asset returns: a new approach*, Econometrica, Vol.59, 347–370, 1991.

[23] D. Bahdanau, K. Cho, and Y. Bengio: *Neural machine translation by jointly learning to align and translate*, CoRR, abs/1409.0473, 2014.

[24] D. Easley, M. Lopez de Prado and M. O'Hara: *Flow toxicity and liquidity in a high frequency world*, Review of Financial Studies, Vol. 25, No. 5, pp. 1457 – 1493, 2012.

[25] D. Easley, M. Lopez de Prado and M. O'Hara: *The volume clock: Insights into the high frequency paradigm*, Journal of Portfolio Management, Vol. 37, No. 2, pp. 118 – 128, 2011.

[26] D. Erdogmus, Y. Rao de Prado and J. Principe: *Recursive Least Squares for an Entropy Regularized MSE Cost Function*, European Symposium on Artificial Neural Networks, pp. 451–456, 2003.

[27] D.E. Rumelhart, and J.L. McClelland: *Parallel distributed processing*, MIT Press, Cambridge, MA, 1986.

[28] D. Miller: *Leveraging BERT for extractive text summarization on lectures*, arXiv:1906.04165, 2019.

[29] D.M.R. Jason, S. Lawrence, T. Jaime, and R.K. David: *Tackling the Poor Assumptions of Naive Bayes Text Classifiers*, In Proceedings of the Twentieth International Conference on Machine Learning, 616–623, 2003.

[30] D.O. Cajueiro and B.M. Tabak: *Ranking efficiency for emerging markets*, Chaos Solitons & Fractals, 22, 349–352, 2004.

[31] D. TjØstheim: *Granger-causality in multiple time series*,Journal of Econometrics Volume 17, Issue 2, Pages 157-176, 1981.

[32] D. Vayanos and J. Wang: *Market Liquidity – Theory and Empirical Evidence,*Handbook of the Economics of Finance Volume 2, Part B, Pages 1289–1361, 2013.

[33] D. Wan, X. Wei and X. Yang: *Liquidity dynamics around intraday price jumps in Chinese stock market,* Journal of Systems Science & Complexity 30, 434–463, 2017.

[34] Downside Hedge: *Twitter indicator for stock market analysis,* www.downsidehedge.com/twitter-indicators/

[35] E. Baek and W. Brock: *A general test for nonlinear Granger causality: bivariate model,* Working paper, Iowa State University, 1992.

[36] E. Collins, S. Ghosh, and C. Scofield: *An application of a multiple neural-network learning system to emulation of mortgage underwriting judgments,*Proc, IEEE Int. Conf. on Neural Networks, 459–466, 1988.

[37] E. F. Fama: *Efficient capital markets: A review of theory and empirical work,* The Journal of Finance, Vol. 25, No. 2,pp. 383-417, 1970.

[38] E. Gilbert and K. Karahalios: *Widespread worry and the stock market,* In Proceedings of the 4th International Conference on Weblogs and Social Media, 58–65, 2010.

[39] E.S. PAGE: *Continuous Inspection Schemes,* Biometrika, 41, pp.100–114, 1954.

[40] E.S. PAGE: *An improvement to Wald's approximation for some properties of sequential tests* ,Journal of Royal Statistics Society B, 16, pp.136–139, 1954.

[41] F. Black: *The price of commodity contracts,* Journal of Financial Economics, 3, 267–179.

[42] F. M. De Bondt Werner and R. Thaler: *Does the stock market overreact?,*The Journal of Finance, 40(3), 793–805, 1985.

[43] F. Marechal, D. Stamate, R. Olaniyan and J. Marek: *On XLE Index Constituents' Social Media Based Sentiment Informing the Index Trend and Volatility Prediction,*

Proceedings of the 10th Intl. Conference on Computatational Collective Intelligence (ICCCI), Springer, LNCS, 2018.

[44] F. Yoav and E.S. Robert: *A decision-theoretic generalization of on-line learning and an application to boosting*, Journal of Computer and System Science, 49, 119–139, 1997.

[45] G. Qiu, X. He, F. Zhang, Y. Shi, J. Bu and C. Chen: *DASA:Dissatisfaction-oriented Advertising based on Sentiment Analysis*, Expert Systems with Applications,37(9):6182–6191, 2010.

[46] G. Ranco, I. Bordino, G. Bormetti, G. Caldarelli, F. Lillo and M. Treccani: *Coupling news sentiment with web browsing data predicts intra-day stock prices*, e-print arXiv:1412.3948, 2014.

[47] G. Sanford and J.E. Stiglitz: *Information and competitive price systems*, American Economic Review, 66, 246–253, 1976.

[48] H. Hotelling: *Analysis of a complex of statistical variables into principal components*, Journal of Educational Psychology, 24:417–441, 1933.

[49] H. Mao, A. Counts and J. Bollen: *Predicting financial markets: comparing survey, news, twitter and search engine data*, arXiv:1112.1051, 2011.

[50] http://www1.fee.uva.nl/cendef/whoiswho/showHP/default.asp?selected=40&pid=6.

[51] https://www.ft.com/content/7b354ff8-1d73-11e8-aaca-4574d7dabfb6.

[52] https://www.investopedia.com/terms/e/emini.asp.

[53] http://www.kibot.com/.

[54] https://www.marketwatch.com/investing/stock/x/charts.

[55] https://www.quandl.com/data/AOS-Alpha–One–Sentiment–Data.

[56] http://research.economics.unsw.edu.au/vpanchenko#software.

[57] https://www.irishtimes.com/business/personal–finance/guesswork–masquerades–as–expertise–in–stock–market–forecasts–1.3358919.

[58] https://uk.finance.yahoo.com/.

[59] https://www.ft.com/content/b9e24c34–bd5b–11e6–8b45–b8b81dd5d080.

[60] https://intrinio.com/.

[61] https://gluebenchmark.com/tasks.

[62] https://www.researchgate.net/publication/251231364_FinancialPhraseBank–v10.

[63] https://school.stockcharts.com/doku.php?id=technical_indicators:chaiki_money_flow_cmf.

[64] https://en.wikipedia.org/wiki/List_of_S%26P_500_companies.

[65] I. Gale and J.E. Stiglitz: *A simple proof that futures markets are almost always informationally inefficient*, Working Paper No. 3209, National Bureau of Economic Research,1050 Massachusetts Avenue Cambridge, MA 02138, 1989.

[66] J. Bollen, H. Mao and X. Zeng: *Twitter mood predicts the stock market*, Journal of Computational Science, 2(1), 1–8, 2011.

[67] J.C. Brada, H. Ernst and J.V. Tassel: *Letter to the Editor—The Distribution of Stock Price Differences: GaussianAfter All?*, Operations Research 14(2):334–340, 1966.

[68] J. Decety and P.L. Jackson: *The functional architecture of human empathy*, Behavioral and Cognitive Neuroscience Reviews, 3, 71–100, 2004.

[69] J. Devlin, M.W. Chang, K. Lee, and K. Toutanova: *Bert: Pre-training of deep bidirectional transformers for language understanding*, arXiv:1810.04805, 2018.

[70] J. Gehring, M. Auli, D. Grangier, D. Yarats and Y.N. Dauphin: *Convolutional sequence to sequence learning*, arXiv:1705.03122v2, 2017.

[71] J.J. Murphy: *Technical Analysis of the Financial Markets: A Comprehensive Guide to Trading Methods and Applications*, New York, NY, USA, Penguin, 1999.

[72] J. Hosking: *Fractional differencing*,Biometrika, Vol. 68, No.1, pp. 165–176, 1981.

[73] J. Howard and S. Ruder: *Universal Language Model Fine-tuning for Text Classification*,arXiv:1801.06146 http://arxiv.org/abs/1801.06146, 2018.

[74] J. Li, A. Sun, J. Han, C. Li: *A Survey on Deep Learning for Named Entity Recognition*,arXiv:1812.09449, 2018.

[75] J.W.J. Wilder: *New Concepts in Technical Trading Systems*, Winston-Salem, 1978.

[76] K. Balog, M. Gilad and R. Maarten de: *Why are they excited? Identifying and explaining spikes in blog mood levels*, In Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL), 2006.

[77] K.A.J. Doherty, R.G. Adams, N. Davey and W. Pensuwon: *Hierarchical Topological Clustering Learns Stock Market Sectors*, ICSC Congress on Computational Intelligence Methods and Applications 1–6, Istanbul, 2005.

[78] K. Boudt, C. Croux and S. Laurent: *Robust estimation of intraweek periodicity in volatility and jump detection*, Journal of Empirical Finance 18, 353–367, 2011.

[79] K. Boudt and M. Pertitjean: *Intraday liquidity dynamics and news releases around price jumps: Evidence from the DJIA stocks*, Journal of Financial Markets 17, 121–149, 2014.

[80] K.J. Astrom and B. Bernhardsson: *Comparison of periodic and event based sampling for first-order stochastic systems 1*, IFAC Proceedings Volumes Vol. 32, Issue 2, Pages 5006-5011 , 1999.

[81] K. Lam and H. Yam: *CUSUM techniques for technical trading in financial markets*, Financial Engineering and the Japanese Markets, Vol. 4, 257 –274, 1997.

[82] K.W. KEMP: *The average run length of a cumulative sum chart when V-mask is used*, Journal of Royal Statistics Society B, 23, pp.149–153, 1961.

[83] K.W. KEMP: *The use of cumulative sums for sampling inspection schemes*, Applied Statistics, 11, pp.16–31, 1962.

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

[85] L. Fortuna, G. Rizzotto, M. Lavorgna, G. Nunnari, M.G. Xibilia and R. Caponetto: *Evolutionary Optimization Algorithms*, In: Soft Computing. Advanced Textbooks in Control and Signal Processing. Springer, London, 2001.

[86] L. Harris: *Trading and electronic markets: what investment professionals need to know*, The Research Foundation of CFA Institute, 2015.

[87] L.V.D. Maaten, E. Postma and J.V.D. Herik: *Dimensionality Reduction: A Comparative Review*, TiCC TR, 2009–05.

[88] M. Baker and J. Wurgler: *Investor sentiment in the stock market*, Journal of Economics Perspectives, 21(2), 129–151.

[89] M. Butler and D. Kazakov: *Particle Swarm Optimization of Bollinger Bands*, Proceedings of the 7th international conference on Swarm intelligence, 2010.

[90] M. Denker and G. Keller: *On U-statistics and von-Mises statistics for weakly dependent processes*, Z. Wahrscheinlichkeitsthorie und Verwandte Gebiete, 64, 505–522, 1983.

[91] M.E. Malliaris, and L. Salchenberger: *A neural network model for estimating option prices*, Journal of Applied Intelligence 3, 193–206, 1993.

[92] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee and L. Zettlemoyer: *Deep contextualized word representations*, 2018 from https://doi.org/10.18653/v1/N18–1202 arXiv:1802.05365.

[93] M. Gevrey, I. Dimopoulos and S. Lek: *Review and comparison of methods to study the contribution of variables in artificial neural network models*, Ecol. Model, 160: 249-264, 2003.

[94] M. Kuhn and K. Johnson: *Applied Predictive Modeling*, Springer, 2013.

[95] M. Maia, S. Handschuh, A. Freitas, B. Davis, R. Mcdermott, M. Zarrouk, A. Balahur and R. Mc-Dermott: *Companion of the web conference 2018 on the web conference 2018*, WWW 2018, Lyon , France, 23–27, 2018.

[96] M. Malliaris and L. Salchenberger: *Using neural networks to forecast the S&P 500 implied volatility*,Neurocomputing 10, 183–195, 1996.

[97] M. Neto, G. Calvalcanti and T. Ren: *Financial Time Series Prediction Using Exogenous Series and Combined Neural Networks*, In Proceedings of International Joint Conference on Neural Networks June, Atlanta, Georgia, 14–19, 2009.

[98] M. Ni and C. Zhang: *An Efficient Implementation of the Backtesting of Trading Strategies*, SpringerVerlag, Berlin, 126–131, 2005.

[99] M. Sewell: *Behavioural finance*, University of Cambridge Retrieved September 7, 2012 from http://www.behaviouralfinance.net/behavioural-finance.pdf.

[100] N. Basalto, R. Bellotti, F. De Carlo, P. Facchi and S. Pascazio: *Clustering stock market companies via chaotic map synchronization*, Physica A, 345 (1–2) ,196–206, 2005.

[101] N. Kalchbrenner, L. Espeholt, K. Simonyan, A.V.D. Oord, A. Graves and K. Kavukcuoglu: *Neural machine translation in linear time*, arXiv preprint arXiv:1610.10099v2, 2017.

[102] N. Maknickienė, A.V. Rutkauskas and A. Maknickas: *Investigation of financial market prediction by recurrent neural network*, Innovative Technologies for Science, Business and Education Vilnius: Vilnius Business College 2(11): 3–8, ISSN 2029–1035, 2011.

[103] N. Oliveira, P. Cortez and N. Areal: *On the predictability of stock market behavior using stocktwits sentiment and posting volume*, Portuguese Conference on Artificial Intelligence, Progress in Artificial Intelligence pp 355–365, 2013

[104] O.E. Barndorff–Nielsen and N. Shephard: *Estimating quadratic variation using realized variance*, Journal of Applied Econometrics 17, 457–477, 2002.

[105] P.D.M. Lopez: *Advances in Financial Machine Learning*, Wiley Publishing ISBN:978–1–119–48208–6, 2018.

[106] P. Huang: *The peasant economy and social change in North China*, Stanford University Press, 1985.

[107] P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala: *Good debt or bad debt: Detecting semantic orientations in economic texts*,Journal of the Association for Information Science and Technology, 65(4):782–796, 2014.

[108] R. Engle: *GARCH 101: The Use of ARCH/GARCH Models in Applied Econometrics*, Journal of Economic Perspectives, Volume 15, Number 4, 157–168, 2001.

[109] R. Olaniyan, D. Stamate, and D. Logofatu: *Social web-based anxiety index's predictive information on S&P 500 revisited*, Proceedings of the 3rd Intl. Symposium on Statistical Learning and Data Sciences, 2015.

[110] R. Olaniyan, D. Stamate, D. Logofatu and L. Ouarbya: *Sentiment and Stock Market Volatility Predictive Modelling - a Hybrid Approach*, Proceedings of the 2nd IEEE/ACM International Conference on Data Science and Advanced Analytics, 2015.

[111] R. Olaniyan, D. Stamate, and I. Pu: *A two-step optimised BERT-based NLP algorithm for extracting sentiment from financial news*, Proc. 17th International Conference on Artificial Intelligence Applications and Innovations (AIAI), Springer, accepted, 2021.

[112] R.G. Clarke, H. De-Silva and S. Thorley: *Fundamentals of futures and options*, The Research Foundation of CFA Institute, 2013.

[113] R.G. Clarke, M.T. FitzGerald, P. Berent and M. Statman: *Market Timing with Imperfect Information*, Financial Analysts Journal, Volume 45, Number 6, 27–36, 1989.

[114] R.J. Shiller: *Irrational Exuberance* Princeton: Princeton University press, 2000.

[115] R.P. Schumaker and H. Chen: *Textual analysis of stock market prediction using breaking financial news: The AZFin text system*, ACM Transactions on Information Systems,27(2), 12:1–19, 2009.

[116] R. Merton: *Option pricing when underlying stock returns are discontinuous*, Journal of Financial Economics 3, 125–144, 1976.

[117] S.A. Monfared, and D. Enke: *Volatility forecasting using a hybrid GJR-GARCH neural network model*, Conference organized by Missouri University of Science and Technology, Philadelphia, 2014.

[118] S. Deng, and A. Sakurai: *Foreign exchange trading rules using a single technical indicator from multiple timeframes*, Proceedings of the 27th International Conference on Advanced Information Networking and Applications Workshops,Barcelona, Spain. IEEE; pp. 207–212, 2013.

[119] S. Kandel and R.F. Stambaugh, *On the Predictability of Stock Returns: An Asset Allocation Perspective*, Journal of Finance 51, 385–424, 1996.

[120] S. Krishnamoorthy: *Sentiment analysis of financial news articles using performance indicators*, Knowledge and Information Systems, 56(2):373–394, 2018.

[121] S.S. Lee and P.A. Mykland: *Jumps in financial markets: A new non–parametric test and jump dynamics*, Review of Financial Studies 21, 2535–2563, 2008.

[122] S.Y. Abu–Mostafa and and A.F. Atiya: *Introduction to financial forecasting*, Applied Intelligence 6, 205–13, 1996.

[123] S. Yanlin, H. Kin-Yip, and L. Wai-man, *Public information arrival and stock return volatility: Evidence from news sentiment and Markov Regime-Switching approach*, Research of Finance, Actuarial Studies and Applied Statistics, The Australian National University, ACT 0200, Australia.

[124] T. Bollerslev, T.H. Law and G. Tauchen: *Risk, jumps, and diversification*, Journal of Econometrics 144, 234–256, 2008.

[125] T. Ding, V. Fang and D. Zuo: *Stock Market Prediction based on Time Series Data and Market Sentiment,* https://pdfs.semanticscholar.org/2c91/447c35fe2d4426b6661b8c8c97f439f3172e.pdf.

[126] T. Fawcett, *An introduction to ROC analysis,* Pattern Recognition Letters 27, 861–874, 2006.

[127] T. G. Anderson: *Return volatility and trading volume: an information flow interpretation of stochastic volatility,* Journal of Finance, 51, 169–204, 1996.

[128] T.G. Abdersen and T. Bollersleve: *Answering the skeptics: yes, standard volatility models do provide accurate forecasts,* International Economic Review 39, 885–905, 1998.

[129] T. Mizuno, T. Ohnishi and T. Watanabe: *Novel and topical business news and their impact on stock market activities,*arXiv:1507.06477v1, 2015.

[130] T.O. Sprenger, A. Tumasjan, P.G. Sandner, and I.M. Welpe: *Tweets and trades: the information content of stock microblogs,* European Financial Management, 20(5), 926–957, 2014.

[131] W.B. Dolan and C. Brockett: *Automatically constructing a corpus of sentential paraphrases,*In Proceedings of the Third International Workshop on Paraphrasing (IWP2005), 2005.

[132] W. Huang, Y. Nakamori and S.Y. Wang: *Forecasting stock market movement direction with support vector machine,* Computers & Operations Research, 32, pp. 2513–2522005, 2005.

[133] W.Y. Lee, C.X. Jiang, and D.C. Indro: *Stock market volatility, excess returns, and the role of investment sentiment,* Journal of Banking and Finance, 26, 2277–2299, 2002.

[134] Y. Kim, C. Denton, L. Hoang, and A.M. Rush: *Structured attention networks,* In International Conference on Learning Representations, 2017.

[135] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov: *RoBERTa: a robustly optimized BERT pretraining approach,* arXiv:1907.11692, 2019.

[136] Y. Liu, Z. Li, H. Xiong, X. Gao and J. Wu, J: *Understanding of internal clustering validation measures*, ICDM , 911–916, 2010.

[137] Y. Wang and R. Gu: *Analysis of efficiency for Shenzhen stock market based on multifractal detrended fluctuation analysis*, International Review of Financial Analysis, 18, 271–276, 2009.

[138] Y. Zhang: *Support vector machine classification algorithm and its application*, International Conference on Information Computing and Applications (ICICA), pp 179–186, 2012.

[139] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler: *Aligning books and movies: Towards story-like visual explanations by watching movies and reading books*, In Proceedings of the IEEE international conference on computer vision, pages 19–27, 2015.

[140] Z. Xiao and D. Enke: *Forecasting daily stock market return using dimensionality reduction*, Expert Systems with Applications 67, 126–39, 2017.

Listing of figures

1.1.1 The former US president announced on the 1st of March, 2018, that he would impose tariff on the steel import. This shows the impact of the proposed tariff on United States Steel Corp. After the announcement there was downward movements in the stock price and stock volume. The picture is taken from [54]. 10

1.2.1 Structure of the thesis. Based on how the chapters relate to each other. 14

3.2.1 Regression model. It presents the information about the fitted volatility line in red and the optimal line in black. The figures in the first column represent plots of NN models with the dataset without sentiment variables. The figures in the second column represent volatility plots of the dataset with sentiment variables. 49

3.2.2 Relative importance. It measures the relative importance of the predictors in the model. Variables on the horizontal lines are the predictors. The variables with values below 0 have a negative relationship with the response variable and those with values above 0 have a positive relationship with the response variable. The response variable is the future volatility. 50

3.2.3 Sensitivity analysis plots. It depicts the forms of the relationship between each explanatory variable with regard to the dependent variable while keeping the other explanatory variables constant. $N1$ and $N2$ denote first and second lagged negative sentiment variables respectively. $P1$ and $P2$ denote first and second lagged positive sentiment variables respectively. $Q1$ and $Q2$ are first and third lagged volatility variables. The dataset is normalised to be between 0 and 1. 52

4.3.1 Data pre-processing process flow that details the processes followed to tackle the complexity of high dimensionality dataset. The three gray boxes represent the three main datasets. Detailed results of the K-means cluster are presented in Appendixes A.0.4 and A.0.5 for the constituents' stock prices and sentiments respectively. 61

4.6.1 The three investment portfolios are presented on two separate charts, each related to Jordan Neural Networks (Jordan NN) at the top and Elman Neural Networks (Elman NN) at the bottom. The values on the y-axis denote portfolio values (£) with an initial value of £5,000. The trends in blue and yellow present the optimised models from the evolutionary optimisation algorithms and ordinary Neural Networks active investment portfolios respectively. The trend in gray represents the passive investment portfolio. 70

6.3.1 The figure depicts the relationships among price, average price, upper and lower bands. ρ is assigned with a value of 1.5. 103

6.3.2 The figure depicts the relationships among pivot point, price, resistance and support. 105

6.5.1 The variable importance representation of the selected variables. 115

7.2.1 Graphical representation of the relationship between sentiments and the S&P 500 E-mini prices. 120
7.3.1 This information captures the datasets explored. The stock market variables and the technical analysis indicators are both extracted from the E-mini S &P 500 stock data. The sentiment dataset is derived from the financial news related to the constituents of the S &P 500. The experimental results from the Financial PhraseBank are obtained from [62] with more detail in chapter 5 . . . 122
7.4.1 Feature importance of the variables explored in the predictive model development. 128

8.3.1 Proposed OXGboost Model Architecture. 138

A.0.1 This was an extract from SNIPPET 3.7 of [105] but with some modification that assumes that labels should be assigned immediately after events. 170
A.0.2 This shows the list of the S&P 500 constituents with closing stock prices that we explored in Chapter 4. Those excluded in the analysis do not have sufficient data, hence their exclusion. Please see [64] for more information about the S&P 500 constituents. . . 171
A.0.3 This shows the list of the S&P 500 constituents with sentiment polarities that we explored in Chapter 4. Those excluded in the analysis do not have sufficient data, hence their exclusion. Please see [64] for more information about the S&P 500 constituents. 172
A.0.4 Cluster pre-processing. This presents the results of the K-means clustering applied to the closing prices of the S&P 500 constituents detailed in Fig. 4.3.1. There are four clusters. 173
A.0.5 Cluster pre-processing. This presents the results of the K-means clustering applied to the sentiments of the S&P 500 constituents detailed in Fig. 4.3.1. There are four clusters. 174

A.0.6 Sample of the results of our proposed optimised BERT-based NLP algorithm: the sentiments extracted from the financial news related to the former US president, Donald Trump. This confirms that the optimised BERT-based NLP model has a high level of accuracy. 175
A.0.7 The sub function required in A.0.8 for observation labelling. . . 175
A.0.8 Proposed Extreme Forward Labelling. 176

A
Appendix

```python
def getBins(events, close):
    '''
    Compute event's outcome (including side information, if provided).
    events is a DataFrame where:
    -events.index is event's starttime
    -events['t1'] is event's endtime
    -events['trgt'] is event's target
    -events['side'] (optional) implies the algo's position side
    Case 1: ('side' not in events): bin in (-1,1) <-label by price action
    Case 2: ('side' in events): bin in (0,1) <-label by pnl (meta-labeling)
    '''
    #1) prices aligned with events
    close_next = close.shift(-1).dropna()
    events_=events.dropna(subset=['t1'])
    px=events_.index.union(events_['t1'].values).drop_duplicates()
    px1=close.reindex(px,method='bfill')
    px2=close_next.reindex(px,method='bfill')
    #2) create out object
    out=pd.DataFrame(index=events_.index)
    out['ret']=px1.loc[events_['t1'].values].values/px1.loc[events_.index]-1
    if 'side' in events_:out['ret']*=events_['side'] # meta-labeling
    out['bin']=np.sign(out['ret'])
    if 'side' in events_:out.loc[out['ret']<=0,'bin']=0 # meta-labeling
    out['ret_next']=px2.loc[events_['t1'].values].values/px1.loc[events_.index]-1
    if 'side' in events_:out['ret_next']*=events_['side'] # meta-labeling
    out['bin_next']=np.sign(out['ret_next'])
    if 'side' in events_:out.loc[out['ret_next']<=0,'bin_next']=0 # meta-labeling
    return out
```

Figure A.0.1: This was an extract from SNIPPET 3.7 of [105] but with some modification that assumes that labels should be assigned immediately after events.

A,AA,AAL,AAP,AAPL,ABBV,ABC,ABT,ACE,ACN,ADBE,ADI,ADM,ADP,ADS,
ADSK,ADT,AEE,AEP,AES,AET,AFL,AGN,AIG,AIV,AIZ,AKAM,ALL,ALTR,
ALXN,AMAT,AME,AMG,AMGN,AMP,AMT,AMZN,AN,ANTM,AON,APA,APC,APD,
APH,ARG,ATVI,AVB,AVGO,AVY,AXP,AZO,BA,BAC,BAX,BBBY,BBT,BBY,BCR,
BDX,BEN,BF-B,BHI,BIIB,BK,BLK,BLL,BMY,BRCM,BRK-B,BSX,BWA,BXP,C,
CA,CAG,CAH,CAM,CAT,CB,CBG,CBS,CCE,CCI,CCL,CELG,CERN,CF,CHK,CHRW,
CI,CINF,CL,CLX,CMA,CMCSA,CMCSK,CME,CMG,CMI,CMS,CNP,CNX,COF,COG,
COH,COL,COP,COST,CPB,CRM,CSC,CSCO,CSX,CTAS,CTL,CTSH,CTXS,CVC,CVS,
CVX,D,DAL,DD,DE,DFS,DG,DGX,DHI,DHR,DIS,DISCA,DISCK,DLPH,DLTR,DNB,
DO,DOW,DPS,DRI,DTE,DUK,DVA,DVN,EA,EBAY,ECL,ED,EFX,EIX,EL,EMC,EMN,
EMR,ENDP,EOG,EQIX,EQR,EQT,ES,ESRX,ESS,ESV,ETFC,ETN,ETR,EW,EXC,EXPD,
EXPE,F,FAST,FB,FCX,FDX,FE,FFIV,FIS,FISV,FITB,FLIR,FLR,FLS,FMC,FOSL,
FOX,FOXA,FSLR,FTI,FTR,GAS,GD,GE,GGP,GILD,GIS,GLW,GM,GMCR,GME,GNW,GOOG,
GOOGL,GPC,GPS,GRMN,GS,GT,GWW,HAL,HAR,HAS,HBAN,HBI,HCA,HCN,HCP,HD,HES,
HIG,HOG,HON,HOT,HP,HPQ,HRB,HRL,HRS,HSIC,HST,HSY,HUM,IBM,ICE,IFF,INTC,
INTU,IP,IPG,IR,IRM,ISRG,ITW,IVZ,JBHT,JCI,JEC,JNJ,JNPR,JPM,JWN,K,KEY,
KIM,KLAC,KMB,KMI,KMX,KO,KORS,KR,KSS,KSU,L,LB,LEG,LEN,LH,LLL,LLTC,LLY,
LM,LMT,LNC,LOW,LRCX,LUK,LUV,LVLT,LYB,M,MA,MAC,MAR,MAS,MAT,MCD,MCHP,MCK,
MCO,MDLZ,MDT,MET,MHFI,MHK,MJN,MKC,MLM,MMC,MMM,MNST,MO,MON,MOS,MPC,MRK,
MRO,MS,MSFT,MSI,MTB,MU,MUR,MYL,NBL,NDAQ,NEE,NEM,NFLX,NFX,NI,NKE,NLSN,NOC,
NOV,NRG,NSC,NTAP,NTRS,NUE,NVDA,NWL,O,OI,OKE,OMC,ORCL,ORLY,OXY,PAYX,PBCT,
PBI,PCAR,PCG,PCL,PCLN,PCP,PDCO,PEG,PEP,PFE,PFG,PG,PGR,PH,PHM,PKI,PLD,PM,
PNC,PNR,PNW,POM,PPG,PPL,PRGO,PRU,PSA,PSX,PVH,PWR,PX,PXD,QCOM,R,RAI,RCL,
REGN,RF,RHI,RHT,RIG,RL,ROK,ROP,ROST,RRC,RSG,RTN,SBUX,SCG,SCHW,SE,SEE,SHW,
SIG,SJM,SLB,SLG,SNA,SNDK,SNI,SO,SPG,SPLS,SRCL,SRE,STI,STJ,STT,STX,STZ,SWK,
SWKS,SWN,SYK,SYMC,SYY,T,TAP,TDC,TE,TEL,TGNA,TGT,THC,TIF,TJX,TMK,TMO,TRIP,
TROW,TRV,TSCO,TSN,TSO,TSS,TWC,TWX,TXN,TXT,TYC,UA,UAL,UHS,UNH,UNM,UNP,UPS,
URBN,URI,USB,UTX,V,VAR,VFC,VIAB,VLO,VMC,VNO,VRSK,VRSN,VRTX,VTR,VZ,WAT,WBA,
WDC,WEC,WFC,WFM,WHR,WM,WMB,WMT,WU,WY,WYN,WYNN,XEC,XEL,XL,XLNX,XOM,XRAY,XRX,
XYL,YHOO,YUM,ZBH,ZION

Figure A.0.2: This shows the list of the S&P 500 constituents with closing stock prices that we explored in Chapter 4. Those excluded in the analysis do not have sufficient data, hence their exclusion. Please see [64] for more information about the S&P 500 constituents.

AA,AAP,AAPL,ABBV,ABC,ABT,ACN,ADBE,ADI,ADM,ADP,ADSK,ADT,AEE,AEP,AES,AET,AFL,
AGLNY,AGN,AIG,ALL,ALXN,AMAT,AME,AMGN,AMP,AMT,AMZN,AN,ANTM,AON,APA,APC,APD,
APH,ARG,ATVI,AVB,AVY,AXP,AZO,BA,BAC,BAX,BBBY,BBT,BBY,BCR,BDX,BEN,BHI,BIIB,
BK,BLK,BLL,BMY,BRKB,BSX,BWA,CAT,CB,CBSA,CCL,CELGZ,CHD,CHK,CI,CMA,CMCSA,CMI,
CNP,COF,COG,COL,COP,COST,CPB,CRM,CSX,CTL,CVS,CVX,D,DAL,DD,DE,DGX,DHR,DIS,
DLTR,DO,DOV,DOW,DRI,DTE,DUK,DVN,EA,EBAY,ECL,ED,EFX,EIX,EL,EMN,EMR,EOG,EQR,
EQT,ESRX,ESS,ESV,ETN,EXC,F,FAST,FB,FCX,FDX,FE,FFIV,FISV,FITB,FLIR,FLR,FLS,
FMC,FOXA,FSLR,GAS,GD,GE,GILD,GIS,GLW,GM,GOOG,GPC,GPS,GRMN,GS,GT,GWW,HAL,HAS,
HBAN,HCA,HCN,HCP,HD,HES,HIG,HOG,HON,HOT,HP,HPQ,HRL,HSY,HUM,IBM,ICE,ILMN,INTC,
INTU,IR,ITW,IVR,JCI,JNJ,JNPR,JPM,JWN,K,KEY,KIM,KLAC,KMB,KMI,KORS,KR,KSS,KSU,
LB,LEG,LENB,LLTC,LLY,LM,LMT,LNC,LOW,LRCX,LUK,LUV,LVLT,LYB,M,MA,MAC,MAR,MAS,
MAT,MCD,MCHP,MCK,MCO,MDLZ,MDT,MET,MHFI,MKC,MLM,MMC,MMM,MNK,MNST,MO,MON,MOS,
MPC,MRK,MRO,MS,MSFT,MSI,MTB,MU,MUR,MYL,NBL,NDAQ,NEE,NEM,NFLX,NFX,NI,NKE,
NLSN,NOC,NOV,NRG,NSC,NTAP,NTRS,NUE,NVDA,NWL,NWS,NWSA,O,OI,OKE,OMC,ORCL,ORLY,
OXY,PAYX,PBI,PCAR,PCG,PCLN,PDCO,PEP,PFE,PFG,PG,PGR,PH,PHM,PKI,PLD,PM,PNC,PNR,
PNW,POM,PPG,PPL,PRGO,PRU,PSA,PSX,PVH,PX,PXD,QCOM,R,RAI,RCL,REGN,RF,RHI,RHT,
RIG,RL,ROK,ROP,ROST,RRC,RSG,RTN,SBUX,SCG,SCHW,SE,SEE,SHW,SIG,SJM,SLB,SLG,SNA,
SNDK,SNI,SO,SPG,SPLS,SRCL,SRE,STI,STJ,STT,STX,SWK,SWKS,SWN,SYK,SYMC,SYY,T,TAP,
TDC,TE,TEL,TGT,THC,TIF,TJX,TMK,TMO,TRIP,TROW,TRV,TSCO,TSN,TSO,TSS,TWC,TWX,TXN,
TXT,TYC,UA,UNH,UNM,UNP,UPS,URBN,URI,USB,UTX,V,VAR,VFC,VIAB,VLO,VMC,VNO,VRSK,
VRSN,VRTX,VTR,VZ,WDC,WEC,WFC,WFM,WHR,WM,WMB,WMT,WU,WY,WYN,WYNN,XEL,XL,XLNX,
XOM,XRX,XYL,YHOO,YUM,ZION,ZTS

Figure A.0.3: This shows the list of the S&P 500 constituents with sentiment polarities that we explored in Chapter 4. Those excluded in the analysis do not have sufficient data, hence their exclusion. Please see [64] for more information about the S&P 500 constituents.

```
CLUSTER 1 - 33 VARIABLES
AA,  APC, APA, BHI, COG, CAM, CF,  CHK, CNX, DVN, DO,  EQT,
FTI, FCX, GNW, HP,  KMI, LYB, MOS, MUR, NOV, NFX, NEM, NRG,
OXY, OKE, PXD, RRC, SWN, SE,  RIG, WMB

CLUSTER 2 - 135 VARIABLES
MMM,   ACN,   ADT,   AFL,   AMG,   APD,   ALL,   AIG,   AME,   APH,
AON,   ADM,   AIZ,   AVY,   BLL,   BA,    BWA,   CHRW,  CA,    COF,
CAT,   CBG,   CBS,   SCHW,  CTAS,  CME,   GLW,   CMI,   DHR,   DE,
DLPH,  DFS,   DISCK, DOW,   DD,    DNB,   ETFC,  EMN,   ETN,   ECL,
EMC,   EMR,   EFX,   EXPD,  FAST,  FDX,   FIS,   FLS,   FMC,   F,
BEN,   GD,    GE,    GM,    GPC,   GT,    GWW,   HRS,   HIG,   HST,
ITW,   IR,    IBM,   IP,    IPG,   IFF,   JEC,   JBHT,  JCI,   KSU,
LLL,   LM,    LEG,   LUK,   LMT,   L,     MLM,   MAS,   MHFI,  MCHP,
MHK,   MON,   MCO,   NDAQ,  NWL,   NLSN,  NSC,   NTRS,  NOC,   NUE,
OMC,   OI,    PCAR,  PH,    PAYX,  PNR,   PBCT,  PPG,   PX,    PCP,
PGR,   PWR,   RTN,   RSG,   RHI,   ROK,   COL,   ROP,   R,     SNI,
SEE,   SHW,   SNA,   SWK,   HOT,   TEL,   TXT,   TRV,   TWX,   TSS,
TYC,   UNP,   UPS,   URI,   UTX,   UNM,   DIS,   WM,    WU,    WHR

CLUSTER 3 - 172 VARIABLES
ABT,   ABBV,  ATVI,  ADBE,  AAP,   A,     ARG,   AKAM,  AGN,   ALXN,
ADS,   GOOG,  ALTR,  AMZN,  AAL,   AXP,   ABC,   AMGN,  ADI,   AAPL,
AMAT,  ADSK,  AN,    AZO,   AVGO,  BCR,   BAX,   BDX,   BBBY,  BBY,
BIIB,  HRB,   BSX,   BMY,   BRCM,  CVC,   CAH,   HSIC,  KMX,   CCL,
CELG,  CERN,  CMG,   CI,    CSCO,  CTXS,  COH,   CTSH,  CMCSK, CSC,
STZ,   DHI,   DRI,   DVA,   XRAY,  DG,    DLTR,  EBAY,  EW,    EA,
ENDP,  EQIX,  EXPE,  ESRX,  FFIV,  FB,    FSLR,  FLIR,  FOSL,  FTR,
GME,   GPS,   GRMN,  GILD,  HBI,   HOG,   HAR,   HAS,   HCA,   HPQ,
HD,    HUM,   INTC,  ICE,   INTU,  ISRG,  JNJ,   JNPR,  GMCR,  KLAC,
KSS,   LB,    LH,    LRCX,  LVLT,  LLY,   LOW,   M,     MAT,   MCK,
MJN,   MDT,   MRK,   KORS,  MU,    MSFT,  MNST,  MSI,   MYL,   NTAP,
NFLX,  NKE,   JWN,   NVDA,  ORLY,  ORCL,  PDCO,  PKI,   PRGO,  PFE,
PBI,   RL,    PCLN,  PVH,   QCOM,  DGX,   RHT,   REGN,  ROST,  CRM,
SNDK,  STX,   SIG,   SWKS,  LUV,   STJ,   SPLS,  SBUX,  SYK,   SYMC,
TGNA,  THC,   TDC,   TSO,   TMO,   TIF,   TWC,   TJX,   TSCO,  TRIP,
FOXA,  UA,    UAL,   URBN,  VFC,   VLO,   VAR,   VRSN,  VRSK,  VRTX,
VIAB,  V,     WBA,   WAT,   ANTM,  WDC,   WFM,   WYNN,  XLNX,  YHOO,
YUM,   ZBH

CLUSTER 4 - 55 VARIABLES
AES,  GAS,  AMT,  T,    AVB,  BF.B, CPB,  CNP,  CTL,  CLX,  KO,
CCE,  CL,   CAG,  ED,   COST, CCI,  CVS,  DPS,  EIX,  ETR,  EL,
EXC,  FE,   GGP,  GIS,  HCP,  HRL,  IRM,  K,    KMB,  KR,   MAC,
MKC,  MCD,  TAP,  MDLZ, POM,  PEP,  PCG,  PM,   PCL,  PG,   O,
RAI,  SJM,  SRCL, SYY,  TGT,  TE,   HSY,  TSN,  VZ,   WMT,  WY
```

Figure A.0.4: Cluster pre-processing. This presents the results of the K-means clustering applied to the closing prices of the S&P 500 constituents detailed in Fig. 4.3.1. There are four clusters.

```
CLUSTER 1 - 7 VARIABLES
BBT, D, DE, DHR, EL, HCP, NVDA

CLUSTER 2 - 8 VARIABLES
ADSK, ALL,  BAX,  EOG,  GLW,  HAL,  HES,  VIAB

CLUSTER 3 - 48 VARIABLES
AA, AES,  AGLNY, AMGN,  AMT,  AXP,  BHI,   BK, CBSA, CHD,  CMCSA,
CTL,  ESRX,  FDX,   FISV,  FLS,   GILD,  HBAN,  INTC,  IR, JNJ,  KEY,
KMI,  KSS,  LLY,   LVLT,  MCO,   MHFI,  MO, MRK,   MYL,   PDCO,  PFE,
PKI,  PVH,  REGN,  SIG,   SLG,  STX,   SWKS,  TRV,   TSO,   TWX,   TXT,
WYNN, XL, XOM,   ZTS

CLUSTER 4 - 79 VARIABLES
AAPL, ABC,  ADBE, ADM,  AEP,  AMAT, AON,  APH,  ATVI, BA,   BCR,  BRKB,
COF,  COP,  DLTR, DRI,  EBAY, EFX,  ETN,  FAST, FLIR, FOXA, GOOG, GRMN,
HOG,  HOT,  HPQ,  HSY,  IBM,  JPM,  LEG,  LMT,  LUK,  M, MAC,  MAT,
MCHP, MNK,  MS,   MTB,  NEE,  NFLX, NLSN, NSC,  O, PAYX, PCAR, PH,
PM,   PNR,  POM,  PSA,  QCOM, RIG,  RTN,  SCHW, SEE,  SNI,  SPLS, SRE,
SYK,  T, TEL,  TIF,  TMK,  TRIP, UNM,  UPS,  URI,  UTX,  VAR,  VMC,
VRSN, VZ,   WFC,  WHR,  WMB,  WU,   YUM
```

Figure A.0.5: Cluster pre-processing. This presents the results of the K-means clustering applied to the sentiments of the S&P 500 constituents detailed in Fig. 4.3.1. There are four clusters.

Date	Ticker	News	Predicted Sentiment
06/02/2019 21:06	XOM	3 Stock Market Winners From TrumpÃ¢Â€Â™s State of the Union	Neutral
06/02/2019 11:08	GE	Winners & Losers from Trump's State of the Union Address	Neutral
24/02/2017 13:16	HPE	Investors wait for Buffett's wisdom, Trump's tax plan and more earnings	Neutral
07/06/2017 22:42	INTU	Cramer: Here's how this anti-Trump software stock has managed to rally	Positive
20/01/2017 19:32	WBA	Stocks Rising as Donald Trump Inherits the White House	Positive
10/11/2016 00:52	KSU	Trump Won: Value Stocks to Buy Now	Neutral
12/07/2018 14:02	WBA	Trump to Meet Lockheed, Mars and Walgreens RepsÃ‚ at U.K. Palace Dinner	Neutral
03/01/2017 17:35	WBA	4 Major Earnings That Could Prop Up the Trump Rally	Positive
10/05/2018 15:41	WFC	Stocks march higher, Trump sets date and time for Kim Jong-Un meeting	Positive
12/10/2018 22:11	ABT	Cramer's game plan: Increased volatility could turn positive with help from Trump, I	Positive
10/02/2017 15:29	JWN	Trickle-Down Ethics at the Trump White House	Neutral
11/11/2016 18:28	ZION	Rallying Super-Regional Banks Now Have 3 Trump Cards In Hand	Positive
07/09/2018 22:13	UPS	Cramer's game plan: Be ready for another Trump tariff-fueled sell-off	Negative
05/07/2019 15:04	WBA	Why Trump is in a mind-blowing sweet spot after the June jobs report	Positive
23/08/2017 15:40	NOC	5 Top Defense Stocks to Buy on Trump's Afghanistan Strategy	Neutral
14/06/2017 00:37	VNO	Trump partner said in running to build FBI headquarters	Positive
16/09/2018 22:40	FDX	Dow Jones Futures Fall: China Trade War Set To Escalate As Trump Tariffs Loom	Negative
07/11/2018 11:51	TWTR	Twitter Trump is overwhelming the economic Trump: Mark Penn	Positive
15/05/2018 17:16	KSS	Are Retail Sales a sign Trump's Tax Cuts are working?	Positive
11/09/2018 23:44	HPE	Trump's Tariff Threat to Hurt Apple: ETFs in Focus	Negative
18/01/2017 17:13	SCHW	[$$] Discount Brokerages to Donald Trump: Keep Tweeting	Neutral
27/09/2018 13:00	HOG	Harley-Davidson Stock Continues to Rise despite Tariff Threats In week 38, Harley-D	Negative
09/02/2017 16:18	WU	Optimism About Trump Tax Plan Lifts Stocks To Record Highs - U.S. Commentary	Positive
06/12/2018 20:45	QCOM	The Latest: Trump greets tech execs at White House	Positive

Figure A.0.6: Sample of the results of our proposed optimised BERT-based NLP algorithm: the sentiments extracted from the financial news related to the former US president, Donald Trump. This confirms that the optimised BERT-based NLP model has a high level of accuracy.

```
def f(row):
    if all([row['label'] == True,row['dummy'] == True, row['posi'] == True]):
        val = 1
    elif all([row['label'] == False, row['dummy'] == True, row['nega'] == True]):
        val = -1
    else:
        val = 0
    return val
```

Figure A.0.7: The sub function required in A.0.8 for observation labelling.

```python
def event_label(dff, column,times,factor = 100,r = 0.2,t = 0.2, freq = '1H'):
    '''
    dff: dataframe that includes the ETF stock price
    column: EFT stock price column name
    times: the number of cumulative return sets e.g. 4 would imply 1*factor,2*factor, 2*factor and 4*factor
    factor: the number of EFT transactions in the initial cumulative sum
    r: Expected minimum return
    t: the threshold factor for identifying events
    freq: the frequency of sampling
    '''
    dff = dff[(dff[column] != 0)]
    df = dff.copy()
    df['price_div'] = df[column].diff().dropna()
    df['datetim'] = df.index
    name = 'sum_'
    exit_time = 'exit_time'
    trend = 'trend_'
    list_names = []
    list2_names = []
    for i in range(times):
        cum_factor = factor * (i +1)
        name_ = name +  str(cum_factor)
        exit_time_ = exit_time +  str(cum_factor)
        df[name_] = df['price_div'][::-1].rolling(window=cum_factor,  min_periods=0).sum()[::-1]
        df[exit_time_] = df['datetim'].shift(-cum_factor)
        list_names.append(name_)
    df['posi'] = (df[list_names] > r).all(1)
    df['nega'] = (df[list_names] < -r).all(1)
    result = df.copy()
    sub_df = result[list_names]
    for j in list(range(1, len(sub_df.iloc[0, :]))):
        trend_ = trend + str(j)
        lss = j - 1
        sub_df[trend_] = sub_df.iloc[:, lss] < sub_df.iloc[:, j]
        list2_names.append(trend_)
    sub_df2 = sub_df[list2_names]
    sub_df2['label'] = sub_df2.eq(sub_df2.iloc[:, 0], axis=0).all(1)
    result['label'] = sub_df2[trend_]
    result['dummy'] = sub_df2['label']
    result['bin_next'] = result.apply(f, axis=1).dropna()
    result = result.resample(freq).first().dropna()
    result = result[(result[column] != 0)]
    result = result.reset_index()
    result.index = result['datetim']
    close = dff[column].copy()
    dailyVol = getDailyVol(close)
    tEvents = getTEvents(close,h=dailyVol.mean() * t )
    eevnt = tEvents.to_frame()
    result = pd.concat([result,eevnt], axis=1).dropna()
    return result
```

Figure A.0.8: Proposed Extreme Forward Labelling.