MACHINE LEARNING APPROACH FOR CRUDE OIL PRICE PREDICTION

A thesis submitted to The University of Manchester for the degree of

Doctor of Philosophy

In the Faculty of Engineering and Physical Sciences

2013

SITI NORBAITI ABDULLAH

School of Computer Science

Table of Contents

Table of Contents 2
List of Figures 7
List of Tables 10
List of Abbreviations 14
Abstract 16
Declaration 17
Copyright Statement 18
About the Author 19
Acknowledgement 20
Chapter 1 Introduction 22
1.1 INTRODUCTION 22
1.2 RESEARCH BACKGROUND 23
1.2.1 The Crude Oil Prices 24
1.2.1.1 The Factors used for Oil Price Determiners 26
1.2.1.2 Factors Affecting the Crude Oil Price Market 27
1.3 RESEARCH SIGNIFICANCE 32
1.3.1 The Domino Effects of Crude Oil Price on Economy 32
1.3.2 Research Objectives 34
1.4 RESEARCH FOCUS 36
1.5 RESEARCH CONTRIBUTION 38
1.6 THESIS STRUCTURE 40
Chapter 2 Crude Oil Price Prediction 42
2.1 CRUDE OIL PRICE MARKET PREDICTION 42
2.1.1 Single Statistical and Econometric Prediction Model 43
2.1.2 Single Artificial Intelligence (AI) Prediction Model 50
2.1.3 Hybrid Prediction Model 54

2.2 CONCLUSION 62

Chapter 3 The Development of a Hierarchical Conceptual Model for the Artificial Neural Network-Quantitative Prediction Model 64

3.1 INTRODUCTION 65

3.2 HIERARCHICAL CONCEPTUAL (HC) MODEL 67

3.2.1 Data Classification Based on Price Turning Points 69

3.2.2 Information Retrieval 73

3.2.3 Feature and Rule Extraction 74

3.3 ARTIFICIAL NEURAL NETWORKS (ANN) FOR THE QUANTITATIVE (ANN-Q) PREDICTION MODEL 78

3.3.1 Objective Determination 80

3.3.2 Data Pre-processing 80

3.3.3 Back Propagation (BP) Trained Neural Network (BPNN) Modelling 82

3.3.4 Network Architecture and Parameters 83

3.3.4.1 The Learning Module 86

3.3.4.2 The Predicting Module 88

3.3.4.3 Performance Evaluation Metrics 89

3.4 SIMULATION TESTS AND RESULTS 91

3.4.1 Sensitivity Analysis: Network Architecture and Parameters 91

3.4.2 A Comparison Analysis of Artificial Neural Network-Quantitative (ANN-Q) Prediction Model with Other Crude Oil Price Prediction Models 98

3.5 CONCLUSION 100

Chapter 4 Content Utilisation and Sentiment Analysis of Data Mined from Google News in Fuzzy Grammar Fragments with the Development of the Fuzzy Expert System for the Rule-based Expert Model 101

4.1 INTRODUCTION 102

4.2 CONTENT UTILISATION 103

4.2.1 Google News-Mining and News Corpus 106

4.2.2 Porter-stemming 108

4.2.3 General Inquirer (GI) Category Selection 110

4.3 TERMINAL GRAMMAR BUILDING 113

4.3.1 Language Structure 114
4.3.2 Context-Free Grammar (CFG) 116
4.3.3 Grammar Parsing 120
4.3.4 Terminal Grammar Definitions 123
4.4 THE APPLICATION OF FUZZY GRAMMAR FRAGMENT EXTRACTION FOR THE FUZZY EXPERT MODEL 126
4.4.1 Text Fragment Selection and Extraction 129
4.4.2 Grammar Fragments Building and Extraction 131
4.5 SENTIMENT MINING AND ANALYSIS WITH RULES EXTRACTION 139
4.6 RULE DELINEATION 141
4.6.1 Rule Delineation with the Decision Tree 141
4.6.2 Simulation Results from the Decision Tree and Other Approaches 147
4.7 FUZZY EXPERT SYSTEM 150
4.7.1 Linguistic Variables and Fuzzy Sets 152
4.7.2 Constructing Fuzzy Rules 156
4.7.3 Building and Evaluating a Fuzzy Expert System 158
4.8 CONCLUSION 162
Chapter 5 Linguistic Prediction Model with Sentiments Mined from Google News Articles and Its Hybridisation with a Quantitative Prediction Model 164
5.1 INTRODUCTION 165
5.2 THE LINGUISTIC INPUT: SENTIMENT MINING FROM THE GOOGLE NEWS ARTICLES 167
5.2.1 Sentiment Mining 167
5.3 THE ARTIFICIAL NEURAL NETWORKS (ANN) LINGUISTIC PREDICTION MODEL 170
5.3.1 Back-Propagation Trained Neural Network (BPNN) Learning Algorithm 171
5.4 THE LINGUISTIC PREDICTION MODEL: ITS NETWORK ARCHITECTURE AND PARAMETER 177
5.4.1 Sensitivity Analysis of Artificial Neural Networks (ANN) and Parameters 180

5.4.2 Sensitivity Analysis: Simulation Test and Results 181

5.4.2.1 Sensitivity Analysis: Results from Normalised Prices as the Training Output with Four and Five Neurons in the Hidden Layer 183

5.4.2.2 Sensitivity Analysis: Results from Normalised Prices as the Training Output with four and five Neurons in the Hidden Layer 186

5.4.2.3 Sensitivity Analysis: Results from the Original Prices as the Training Output with Four and Five Neurons in the Hidden Layer 188

5.4.2.4 Overall Result Evaluation for the Linguistic Prediction Model 190

5.4.2.5 Comparison of the Linguistic Prediction Model Results with other Machine Learning Approaches 191

5.5 THE LINGUISTIC AND QUANTITATIVE (LQ) PREDICTION MODEL: THE HYBRID 194

5.5.1 Back-Propagation Neural Network (BPNN) Application in the Linguistic-Quantitative (LQ) Hybrid Prediction Model 195

5.6 THE LINGUISTIC-QUALITATIVE (LQ) PREDICTION MODEL: ITS NETWORK ARCHITECTURE AND PARAMETERS 199

5.6.1 Sensitivity Analysis: Simulation Test and Results 204

5.6.1.1 Sensitivity Analysis: Results from Training Set A with 4 and 5 Neurons in the Hidden Layer 206

5.6.1.2 Sensitivity Analysis: Results from Training Set B with 4 and 5 Neurons in the Hidden Layer 209

5.6.1.3 Sensitivity Analysis: Results from Training Set C with 4 and 5 Neurons in the Hidden Layer 214

5.6.1.4 Overall Results for the Linguistic-Quantitative (LQ) Prediction Model 220

5.6.1.5 Comparison of the Results with Other Machine-Learning Approaches 222

5.7 CONCLUSION 226

Chapter 6 Conclusion and Suggestions for Future Research 228

6.1 INTRODUCTION 228

6.2 MAIN CONTRIBUTIONS OF THE THESIS 233

6.2.1 Hierarchical Conceptual Model and Artificial Neural Network-Quantitative (ANN-Q) Prediction Model 234

6.2.2 Content Utilisation and Sentiment Analysis with Fuzzy Grammar Fragment Extraction Application for Rule-based Expert Model 235

6.2.3 Linguistic Prediction Model and Linguistic-Quantitative (LQ) Hybrid Model 236

6.3 SUGGESTIONS FOR FURTHER RESEARCH 237

Appendix A Sample of Terminal Grammar ('Dictionary'): The *Header* 253

Appendix B Sample of Fuzzy Inference System Structure 256

Appendix C Linguistic Prediction Model: The Computation 258

A-BEST PREDICTION RESULTS BASED ON TESTING DATA WITH 90:10 RATIO 258

B- BEST PREDICTION RESULTS BASED ON TESTING DATA WITH 80:20 RATIO 260

C- BEST PREDICTION RESULTS BASED ON TESTING DATA WITH 70:30 RATIO 262

Word Count: 51,751

List of Figures

Figure 1.1 The Crude Oil Price Market with the Events Related to its Volatility for January 2007 to October 2011 [9].....23
Figure 1.2 The West Texas Intermediate (WTI) Price (Monthly) 24
Figure 1.3 United States (US) Petroleum Consumption for Year 1973 to 2011 [9]..31
Figure 2.1 The Comparison of Dynamic Forecasts and West Texas Intermediate (WTI) Price [20].....48
Figure 2.2 Support Vector Machine (SVM)-based Forecasting System Procedures.51
Figure 2.3 The TEI@I Methodology Framework for Crude Oil Price Forecasting [41]. 55
Figure 3.1 The Development Framework for Hierarchical Conceptual Model and Quantitative Prediction Model. 66
Figure 3.2 The Development Process of the Hierarchical Conceptual Model.....68
Figure 3.3 Monthly Crude Oil Price Turning Points for Jun 2007 to December 2009. 69
Figure 3.4 The Key Impact Factors Contributing to Crude Oil Price Volatility Based on Hierarchical Conceptual (HC) Model. 75
Figure 3.5 Artificial Neural Networks (ANN) Development Framework [50]. 79
Figure 3.6 The Quantitative Prediction Model Optimal Performance Based on 80:20 Percent Training: testing Data Ratio with Hidden Neurons = 5. 96
Figure 3.7 The Error Mapping of the Quantitative Prediction Model Based on 80:20 Percent Training: testing Data Ratio with Hidden Neurons = 5. 97
Figure 4.1 The Content Utilisation and Sentiment Analysis Framework for Google News.....105
Figure 4.2 Categorisation Process of Lexical Items into Positive and Negative Content Categories. 112
Figure 4.3 An Example of English Language Structure 114
Figure 4.4 Example Of Context-Free Grammar (CFG). 117
Figure 4.5 Examples of Context-Free Grammar (CFG) Rules For the Crude Oil Price.....118
Figure 4.6 Example of a Context-Free Grammar (CFG) Parse Tree for the Crude Oil Price.....119

Figure 4.7 The Fuzzy Expert Model Framework with Text Fragment and Grammar Extraction, Sentiment Mining and Analysis and Rule Extraction Structures. 128
Figure 4.8 Rule Delineation Performed with the J48 Decision Tree for the Rule-Based Expert Model.....148
Figure 4.9 The Development Framework of the Fuzzy Expert System through a Fuzzy Expert Model.....151
Figure 4.10 Example of Normalised Fuzzy Sets for Input *oilFactor*=Demand. 155
Figure 4.11 Example of Normalised Fuzzy Sets for Output=*Price*. 155
Figure 4.12 A Fuzzy Rule Base for Price 157
Figure 4.13 Example of Inference System Output Surface View. 159
Figure 4.14 Example of Rule Viewer for Price.....160
Figure 5.1 Sentiment Mining and Linguistic Prediction Model Framework. 169
Figure 5.2 A Three-Layer Artificial Neural Network (ANN) Trained with Back-Propagation (BP) Topology in a 13-5-1 Structure. 172
Figure 5.3 The Development of the Back-Propagation Neural Network (BPNN) for the Linguistic Prediction Model.....178
Figure 5.4 The Linguistic-Quantitative (LQ) Prediction Model Framework with the Back-Propagation Neural Network (BPNN).....196
Figure 5.5 Three-Layer Artificial Neural Networks (ANN) Trained with Back-Propagation Topology of 35-5-1 for the Linguistic-Quantitative (LQ) Prediction Model 202
Figure 5.6 The Best Performance Result for Training Set A, Subset A-1 of the Linguistic-Quantitative (LQ) Prediction Model with 90:10 Per cent Data Ratio and 4 Hidden Neurons. 207
Figure 5.7 The Best Performance Result for Training Set A, Subset A-2 of the Linguistic-Quantitative (LQ) Prediction Model with 90:10 Per cent Data Ratio and 5 Hidden Neurons. 208
Figure 5.8 The Best Directional Performance Result (90.00%) for Training Set B, Subset B-1 of the Linguistic-Quantitative (LQ) Prediction Model based on the Directional Statistics (D_{stat}) Value, Trained with an 80:20 Per cent Data Ratio and 4 Hidden Neurons. 210
Figure 5.9 The Best Directional Performance Result (90.00%) for Training Set B, Subset B-2 of the Linguistic-Quantitative (LQ) Prediction Model based on the

Directional Statistics (D_{stat}) Value, Trained with an 80:20 Per cent Data Ratio and 5 Hidden Neurons. 211
Figure 5.10 The Best Performance Result for Training Set B of the Linguistic-Quantitative (LQ) Prediction Model with a 70:30 Per cent Data Ratio and 5 Hidden Neurons. 213
Figure 5.11 The Best Directional Performance Result (85.71%) with 0.969 NMSE for Training Set C, Subset C-1 of the Linguistic-Quantitative (LQ) Prediction Model based on the Directional Statistic (D_{stat}) Value, Trained with a 70:30 Per cent Data Ratio and 4 Hidden Neurons. 216
Figure 5.12 The Best Directional Performance Result (80.00%) with 0.938 NMSE for Training Set C, Subset C-2 of the Linguistic-Quantitative (LQ) Prediction Model based on the Directional Statistic (D_{stat}) Value, Trained with an 80:20 Per cent Data Ratio and 5 Hidden Neurons. 217
Figure 5.13 The Error Mapping for the Best Directional Performance Result (85.71%) with 0.969 NMSE for Training Set C, Subset C-1 of the Linguistic-Quantitative (LQ) Prediction Model based on Directional Statistic (D_{stat}) Value, Trained with a 70:30 Percent Data Ratio and 4 Hidden Neurons. 218
Figure 5.14 The Error Mapping for the Best Directional Performance Result (80.00%) for Training Set C, Subset C-2 of the Linguistic-Quantitative (LQ) Prediction Model based on the Directional Statistic (D_{stat}) Value, Trained with an 80:20 Percent Data Ratio and 5 Hidden Neurons. 219
Figure 5.15 The Back Propagation Neural Networks (BPNN) Performance for the Linguistic-Quantitative (LQ) Prediction Model: A Comparison with Other Machine-Learning Approaches Based on Training Sets A-2, B-2 and C-1. 223
Figure 5.16 The Back- Propagation Neural Networks (BPNN) Performance for the Linguistic-Quantitative (LQ) Prediction Model Based on the Minimum RMSE Value: A Comparison with Other Machine-Learning Approaches. 224

List of Tables

Table 1.1 The Top Non-OPEC Oil-producing Countries in the World (2013 est.)...29
Table 2.1....44
Table 2.2....52
Table 2.3 The Root Mean Square Error (RMSE) Result from Simulation Experiments....60
Table 2.4 The Impact Factors used by Different Models for Crude Oil Price Prediction Model....63
Table 3.1 Turning Points Impact Category Based on Data Classification Process 70
Table 3.2 Example on Classification of Price Impact in Excel....72
Table 3.3 The Key Factors Influencing Crude Oil Price and used for the Quantitative Prediction Model....77
Table 3.4 Example of Data Represented in Normalised Form for the Quantitative Prediction Model....81
Table 3.5 Back Propagation Neural Network (BPNN) Parameters for Quantitative Prediction Model....84
Table 3.6 Example of Artificial Neural Networks (ANN) Architecture....85
Table 3.7 Sensitivity Analysis for the Quantitative Data Employed in Back Propagation Neural Networks (BPNN)....86
Table 3.8 Example of Actual Data vs Target Predicted Data Represented in Time Series with Hidden Neurons = 4....91
Table 3.9 The Absolute Error Results for Training Sets with Hidden Neurons = 3, 4 and 5, and Trained with Normalised 90:10 percent, 80:20 percent and 70:30 percent of Training: testing Data Ratio for the Quantitative Prediction Model. 92
Table 3.10 The Results from Sensitivity Analysis for the Normalised Quantitative Data Employed in Back Propagation Neural Networks (BPNN) with 80:20 Percent Ratio of Total Data and Hidden Neuron = 5....94
Table 3.11 The Performance Results based on the Sensitivity Analysis Made with the Parameters Depicted in Subsection 3.3.4, Trained with Normalised Data and Hidden Neurons = 5....95
Table 4.1 Example of a Porter-stemmed Article (January 2009)....109
Table 4.2 Content Categories Selected from the General Inquirer (GI) 111

Table 4. 3 Grammar Building Structure: Word Categories 115
Table 4.4 Grammar Building Structure: Examples of Constituent Phrases.....115
Table 4.5 Grammar Parsing: Examples of Bottom-Up, Shift-Reduce Parsing.....121
Table 4.6 The Content Definitions of a Terminal Grammar.....124
Table 4.7 Example of Extracted Text Fragments Derived from the Training with Content Size= 5, *coreTerm= OilFactor= Demand* for a News Article extracted in January 2008. 130
Table 4.8 Examples of Extracted Grammar Fragments 134
Table 4.9 Examples of Extracted Grammar Fragments 136
Table 4.10 Examples of Extracted Grammar Fragments 137
Table 4.11 Example of a Database for Analysed Sentiments Derived from Extracted Grammar Fragments with *coreTerm = oilFactor*, October 2009.....142
Table 4.12 The Inputs = *coreTerm = oilFactor* Data used for Decision Tree.....144
Table 4.13 J48 Decision Tree Classification Criteria 145
Table 4.14 The Input *coreTerm = oilFactor* Data Mapped and Established from the Decision Tree with *Price* as the Decision Output.....146
Table 4.15 Results of Rule Delineation and its Comparison with other Machine Learning Approaches. 149
Table 4.16 J48 Decision Tree Confusion Matrix 152
Table 4.17 The Rule Evaluation with the Fuzzy Expert System 161
Table 5.1 The Linguistic Features of the Linguistic Prediction Model. 176
Table 5.2 Back-Propagation Neural Network (BPNN) Parameters.....179
Table 5.3 The Linguistic Prediction Model Results with Training Output= Directional Price, Hidden Neuron= 4. 183
Table 5.4 The Linguistic Prediction Model Results with Training Output= Directional Price, Hidden Neuron= 5. 184
Table 5.5 The Best Linguistic Prediction Model Results for Training Output= Directional Price with Hidden Neuron= 4 and 5.....185
Table 5.6 Linguistic Prediction Model Results with Training Output= Normalised Price, Hidden Neuron= 4. 186
Table 5.7 Linguistic Prediction Model Results with Training Output= Normalised Price, Hidden Neuron= 5. 187
Table 5.8 The Best Linguistic Prediction Model Results for Training Output= Normalised Price with Hidden Neuron= 4 and 5.....187

Table 5.9 Linguistic Prediction Results with 188
Table 5.10 Linguistic Prediction Results with 189
Table 5.11 The Best Linguistic Prediction Model Results for Training Output= Original Price with Hidden Neuron= 4 and 5. 189
Table 5.12 The Best Simulation Results for Linguistic Prediction Model. 190
Table 5.13 The Results of the Linguistic Prediction Model from 193
Table 5.14 The Comparison of the Best Results from the Support Vector Machine (SVM), the Linear Regressions (LR) and the Gaussian Process (GP) Approaches with the Back-Propagation Neural Networks (BPNN) used in the Linguistic Prediction Model. 193
Table 5.15 Data Sets For the Linguistic-Quantitative (LQ) Prediction Model. 198
Table 5.16 Back-Propagation Neural Network (BPNN) Parameters for the Linguistic-Quantitative (LQ) Prediction Model 200
Table 5.17 The Results for the Linguistic-Quantitative (LQ) Prediction Model with 206
Table 5.18 The Results for the Linguistic-Quantitative (LQ) Prediction Model with 206
Table 5.19 The Best Results for the Linguistic-Quantitative (LQ) Prediction Model for 207
Table 5.20 The Results for the Linguistic-Quantitative (LQ) Prediction Model with 209
Table 5.21 The Results for the Linguistic-Quantitative (LQ) Prediction Model with 209
Table 5.22 The Best NMSE Results for the Linguistic-Quantitative (LQ) Prediction Model for 212
Table 5.23 The Results for the Linguistic-Quantitative (LQ) Prediction Model with 214
Table 5.24 The Results for the Linguistic-Quantitative (LQ) Prediction Model with 214
Table 5.25 The Best Performance Results for the Linguistic-Quantitative (LQ) Prediction Model for Training Set C. 215
Table 5.26 The Best Prediction Performance for the Linguistic-Quantitative (LQ) Prediction Model 220
Table 5.27 The Best Prediction Performance of Training Sets A, B, and C for 221

Table 5.28 The Results of the Linguistic-Quantitative (LQ) Prediction Model from 222
Table 5.29 The Comparison of the Best Results from the Support Vector Machine (SVM), the Linear Regressions (LR) and the Gaussian Process (GP) Approaches with.....224
Table 5.30 The Summary of the Optimal Performance for Models in Chapter 5....227
Table 6.1 The Key Impact Factors of Crude Oil Market. 229
Table A-1 Best Prediction Results For Directional Price with Hyperbolic Tangent, Hidden Layer Neuron=5 258
Table A-2 Best Prediction Results For Disnormalised Price with Hyperbolic Tangent, Hidden Layer Neuron=4 258
Table A-3 Best Prediction Results For Original Price with Log Sigmoid, Hidden Layer Neuron=4 259
Table B-1 Best Prediction Results For Directional Price with Hyperbolic Tangent, Hidden Layer Neuron=4 260
Table B-2 Best Prediction Results For Normalised Price with Hyperbolic Tangent, Hidden Layer Neuron=5 260
Table B-3 Best Prediction Results For Original Price with Hyperbolic Tangent, Hidden Layer Neuron=5 261
Table C-1 Best Prediction Results For Directional Price with Hyperbolic Tangent, Hidden Layer Neuron=5 262
Table C-2 Best Prediction Results For Normalised Price with Hyperbolic Tangent, Hidden Layer Neuron=5 262
Table C-3 Best Prediction Results For Original Price with Hyperbolic Tangent, Hidden Layer Neuron=5 263

List of Abbreviations

AI	Artificial Intelligence
ALNN	Adaptive Linear Neural Network
ANN	Artificial Neural Network
ANN-Q	Artificial Neural Networks-Qualitative
ARCO	Atlantic Richfield Company
ARIMA	Autoregressive Integrated Moving Average
ASPO	Association Study of Peak Oil
BN	Belief Network
BPNN	Back Propagation Neural Network
CFG	Context Free Grammar
CPI	Consumer Price Index
D_{stat}	Directional Statistic
ECB	European Central Bank
EIA	Energy Information Administration
EMD	Empirical Mode Decomposition
FNN	Feed-forward Neural Network
FSU	Former Soviet Union
GARCH	Generalised Autoregressive Conditional Heteroskedasticity
GDP	Growth Domestic Products
GI	General Inquirer
GP	Gaussian Processes
GRNN	General Regression of Neural Network
HC	Hierarchical Conceptual
ID	Influence Diagrams
IEA	International Energy Agency
IMF	Intrinsic Mode Functions
KB	Knowledge Base
LQ	Linguistic-Quantitative
LR	Logistic Regression
MALT	Modified Alternative

MAPE	Mean Absolute Percentage Error
MLP	Multi-Layer Perceptron
MMI	Man Machine Interface
NMSE	Normalised Mean Square Error
NYMEX	New York Mercantile Exchange
OPEC	Members of The Organisation of Petroleum Exporting
RBF	Radial Basis Function
RES	Rules-based Expert System
RMSE	Root Mean Square Error
RSK	Relative Stock Model
SMO	Support Vector Regression
STEP	Sentiment Tag Extraction Program
SVM	Support Vector Machine
UK	United Kingdom
US	United States
USC	University of Southern California
USD	US Dollar
WTI	West Texas Intermediate
WTM	Web-based Text Mining

Abstract

Crude oil prices impact the world economy and are thus of interest to economic experts and politicians. Oil price's volatile behaviour, which has moulded today's world economy, society and politics, has motivated and continues to excite researchers for further study. This volatile behaviour is predicted to prompt more new and interesting research challenges. In the present research, machine learning and computational intelligence utilising historical quantitative data, with the linguistic element of online news services, are used to predict crude oil prices via five different models: (1) the Hierarchical Conceptual (HC) model; (2) the Artificial Neural Network-Quantitative (ANN-Q) model; (3) the Linguistic model; (4) the Rule-based Expert model; and, finally, (5) the Hybridisation of Linguistic and Quantitative (LQ) model. First, to understand the behaviour of the crude oil price market, the HC model functions as a platform to retrieve information that explains the behaviour of the market. This is retrieved from Google News articles using the keyword "Crude oil price". Through a systematic approach, price data are classified into categories that explain the crude oil price's level of impact on the market. The price data classification distinguishes crucial behaviour information contained in the articles. These distinguished data features ranked hierarchically according to the level of impact and used as reference to discover the numeric data implemented in model (2). Model (2) is developed to validate the features retrieved in model (1). It introduces the Back Propagation Neural Network (BPNN) technique as an alternative to conventional techniques used for forecasting the crude oil market. The BPNN technique is proven in model (2) to have produced more accurate and competitive results. Likewise, the features retrieved from model (1) are also validated and proven to cause market volatility. In model (3), a more systematic approach is introduced to extract the features from the news corpus. This approach applies a content utilisation technique to news articles and mines news sentiments by applying a fuzzy grammar fragment extraction. To extract the features from the news articles systematically, a domain-customised 'dictionary' containing grammar definitions is built beforehand. These retrieved features are used as the linguistic data to predict the market's behaviour with crude oil price. A decision tree is also produced from this model which hierarchically delineates the events (i.e., the market's rules) that made the market volatile, and later resulted in the production of model (4). Then, model (5) is built to complement the linguistic character performed in model (3) from the numeric prediction model made in model (2). To conclude, the hybridisation of these two models and the integration of models (1) to (5) in this research imitates the execution of crude oil market's regulators in calculating their risk of actions before executing a price hedge in the market, wherein risk calculation is based on the 'facts' (quantitative data) and 'rumours' (linguistic data) collected. The hybridisation of quantitative and linguistic data in this study has shown promising accuracy outcomes, evidenced by the optimum value of directional accuracy and the minimum value of errors obtained.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

About the Author

Siti Norbaiti Abdullah was born in Perak, Malaysia, in 1981. She received a Bachelor of Accounting (Hons.), from the Universiti Tenaga Nasional (UNITEN), Malaysia and her Master of Information Technology (MIT), from The National University of Malaysia in 2003 and 2007, respectively. In September 2008, she joined the University of Manchester, United Kingdom, to pursue her PhD degree.

Acknowledgement

In the name of God (Allah), The Most Beneficent, The Most Merciful.

First of all, I would like to send my utmost gratitude to Allah s.w.t; the Almighty God who has made this journey possible—without His help and His guidance, walking through this journey of success would not have been smooth. I thank Him a lot for always being in every moment of my life, in every inch of my journey.

So verily, with the hardship, there is relief. Verily, with the hardship, there is relief. So when you have finished (from your occupation), then stand up for Allah's worship (prayer). And to your Lord (Alone) turn (all your intentions and hopes) and your invocations[1].

No words can ever describe this gratefulness: to have always under His observation and guidance. Every one's PhD journey is unique. Mine too, without exception. I thank Him for helping me through, for making me strong and for making *this* a very special journey.

None of this would have been possible without the families who have offered constant 'cushions', words of the world, and supports through all this while. I thank my dear husband, *Syah*, for all his kindness, comforts, hugs and love. I thank him so much for his patience and understanding and for giving me the space when I need it the most. (He is now a better cook, which makes me love him more).

Thank you my dear for always being there for me.

To my wonderful, intelligent, thoughtful and two-in-a-million children; well, they knew who they are to me—my teddies, my best buddies; my strengths. I thank *Allah* for bestowing me with two very smart and special children: my lovely, beautiful, and creative girl- *Aufa;* and my cheeky, brilliant, and *soleh* boy- *Amsyar.* I thank them for their constant love. I will never forget the moment when we were comforting each other, counting the blessings of *Allah*, putting the 'strength' and 'content' in our hearts when they both were diagnosed ill. I will never forget when they tip-toed with their plates full of food, prepared for me whenever my health, was not at its best. I

[1] Al-Quran; 94: 5-8.

thank *Allah* for *Aufa* and *Amsyar's* unconditional love. May they grow well and blossom as great individuals—under His constant love and guidance.

Thank you for being my best strength.

I would also like to take this opportunity to thank both of my parents, *Haji Abdullah* and *Hajjah Rohana*, and my parents-in-law, *Mohd Zainuddin* and *Zainah* for always making *du'a* (prayers) for me, for always putting their trust in me, for always loving me. Without their *du'a*s, I would not have reached this far. May *Allah* s.w.t bless all of you and grant you *Jannah* (paradise), *insha Allah* (with God's will).

My utmost gratitude also goes to my supervisor, *Dr Xiao-Jun Zeng* who has guided me through my PhD and also who has been supportive in giving me advice and giving me my momentum back every after meeting. I thank him for giving me space during my hard times in this journey. His humble attitude, commitment, flexibility and kindness are one of a kind. He motivates me to be a better me.

I would like to also thank my sponsor, *Majlis Amanah Rakyat (MARA)*, and the Malaysian government for funding me throughout my years of study.

To my brothers and sister, my families back in Malaysia whom I have not met for a long time, this thank you also goes to all of you.

Last but not least, my friends who are all over the world, your constant prayers and thoughts I will always cherish. May *Allah* bestow you all greatness in here and thereafter, *insha Allah.*

A big thank you too given to both of my examiners, *Dr Richard Neville* and *Dr Keeley Crockett* who are very kind and helpful in giving constructive advice on making the thesis better.

"Thank you, Allah. Thank you, all. Thank you to everyone!"

Chapter 1 Introduction

Overview

This chapter will briefly introduce the research via insight into the research background, its significance, the objectives and focus area, and, finally, its contribution to the knowledge.

1.1 INTRODUCTION

The crude oil price market and its associated high volatility are an interesting area of research. The drastic price increments in the beginning of 2006 shocked the world [1] [2]. Crude oil price is traded on the New York Mercantile Exchange (NYMEX) [3] along with other energy and mineral commodities. As oil is one of the most volatile commodities [4], predicting its trend is computationally complex and intensely problematic. Its volatility is dependent on numerous factors and it is market-sensitive. Due to strong non-determinist effects, market prices change due to many factors. Moreover, its volatility is also a major concern to both oil-importing and oil-exporting countries.

1.2 RESEARCH BACKGROUND

The crude oil price contributes to over 50% of the average price of petroleum [5]. The fluctuations in the crude oil price market mirror petroleum prices. The crude oil price, benchmarked by the West Texas Intermediate (WTI) [6] price in NYMEX [3], has increased drastically since the middle of 2004, with the average price of USD31.14 per barrel[3] in 2003 to the average price of USD56.47 per barrel in 2005. For the period of January to December 2006, the WTI price remained high at an average of USD66.10 per barrel. The price was on an upward curve for the whole year of 2007 and reached its peak on July 2008. The crude oil price reached a record high of USD147.27 on July 11, 2008 [7], and it later declined to a monthly average price of USD39.16 on February 2009 [8]. The volatility of the crude oil prices between 2007 and 2009 was recorded by [9] and presented in Figure 1.1 with related events mapped on the price's turning points.

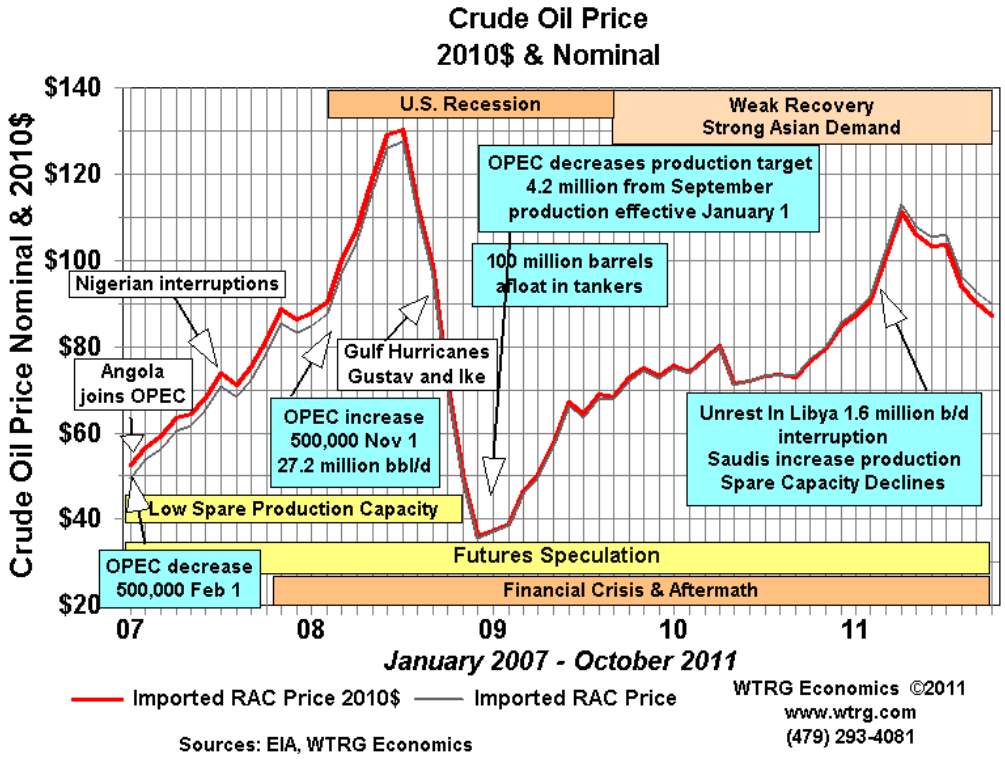


Figure 1.1 The Crude Oil Price Market with the Events Related to its Volatility for January 2007 to October 2011 [9].

[3] Barrel: unit of value for crude oil or petroleum products.

1.2.1 The Crude Oil Prices

Crude oil prices are determined by bidding[5] on oil futures contracts[6] by commodities' traders who are also investors and regulators.

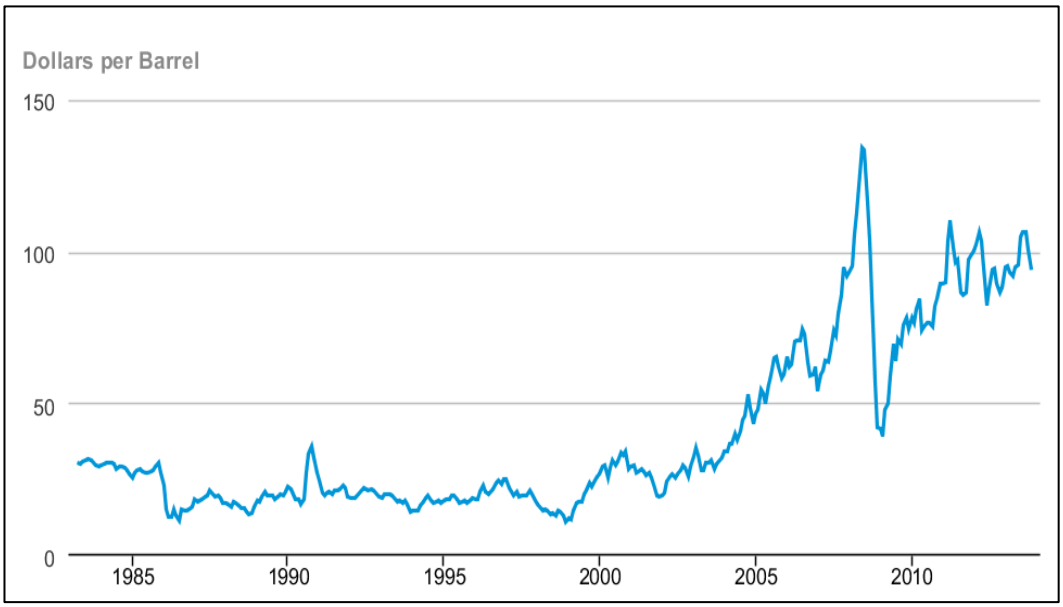


Figure 1.2 The West Texas Intermediate (WTI) Price (Monthly) from 1984 to 2013 [10].

These contracts are agreements between traders to buy or sell oil in the market at a particular date in the future for an agreed-upon price. The execution of oil futures contracts are established through an exchange by persons and firms who are registered with the Commodities Futures Trading Commissions (CFTC)[9] [11]. Figure 1.2 shows the monthly WTI future prices from 1980 to 2013, which also shows the high price hike occurring between 2007 and 2008. The price hike was due to rapidly emerging economies in China and India becoming among the biggest oil consumers in the world.

Commodities traders[10] are categorised by, first, the representatives of oil companies, who actually retail oil. The oil companies purchase oil from the market in order to

[5] Bidding: an offer (often competitive) of setting a price one is willing to pay for something, e.g., oil futures.
[6] Oil future contracts: an agreement to either buy or sell oil on a publicly traded exchange, i.e., NYMEX.
[9] Commodities Futures Trading Commissions (CFTC): an independent agency that regulates futures and options market.
[10] Commodities traders: persons responsible for the execution of the oil futures in a publicly traded exchange.

future deliver the oil to end-users with a fixed price. By locking the oil price into a contract, it enables companies to financially plan to reduce (or hedge[11]) the uncertainty risks and to maintain the financial stability. Good understanding and anticipation of market irregularities are crucial in maintaining the financial stability. The second commodities traders in the oil market are the speculators, who are not buyers or retailers in the market, but bid the market in advance with an anticipated future price. These biddings were made in the NYMEX market through a financial instrument called derivatives[12], wherein value is determined by the value of a commodity: oil price. Thus, these bids in the market by these two traders, who are also the market's regulators, create the oil price.

[11] Hedge: to minimise or offset financial risk. It is also an investment term defined as to prevent or offset another potentially risky or uncontrollable situation.

[12] Derivatives: a financial instruments whose value is based on the performance of underlying assets such as crude oil.

1.2.1.1 The Factors used for Oil Price Determiners

Oil prices in the market are greatly determined by commodities traders' activities [12]. Prices are determined by the way commodities traders observe possible factors related to the market in order to develop a bid. The market factors that draw the attention from traders:

i) The current supply—mainly the current production quota set by the Organisation of Petroleum Exporting Countries (OPEC). If production is presumed to decline in an estimated future date, the traders will bid the oil price up, but they will purchase as much the oil futures to bid the price down if it is presumed that supply will increase in an estimated future.

ii) Oil reserves—stored by the Strategic Petroleum Reserves[17] to be added to the oil supply when prices get too high. A tap on the reserves by one of the oil-reserving countries will lead the traders to bid the price up. Saudi Arabia, as one of the petroleum-exporting countries with a large reserve capacity, is much in the spotlight when a price is hiked.

iii) Oil demand—it is hard to actually anticipate demands. The estimate demands of oil-exporting countries, like the US, are easily accessed from the Energy Information Agency (EIA) [10]. Other demands can be estimated by way of weather forecasts as well as the population of a country. For example, in summer, the weather increases the consumption of petrol by vehicle drivers who head for vacations; likewise, during winter, weather forecasts are used to estimate the consumption of heating oil used in homes. When these factors are examined by traders who are also the market's regulators and opportunists, they will take this opportunity to bid the price up for profit-gaining.

It is known that crude oil prices are not just determined by the expectations and activities of the commodities traders', but they are also affected by other important and related factors like crisis and disasters. Subsection 1.2.1.2 will discuss the different aspects that affect the crude oil market other than traders' anticipation.

[17] Strategic Petroleum Reserves: emergency stockpiles of oil established by member countries of the International Energy Administration to protect against supply disruption.

1.2.1.2 Factors Affecting the Crude Oil Price Market

This subsection will discuss the factors that affect the crude oil price market as an extension to the factors discussed in subsection 1.2.1.1. These factors have fed the steady, sometimes drastic fluctuations of oil prices in recent years. Nelson, et al. [13] has suggested that the major aspects affecting the crude oil market are the demands, supplies, population, geopolitical risks and economic issues:

i) Demand—Global energy demand depends on the population and the economic growth of a country. As incomes rise, economies use more energy for transport, heating and cooling and producing goods and services. China and India, with a combined population of about 2.4 billion [10], began their economic growth in the 1990s with China's per capita[19] Growth Domestic Product (GDP) rising from USD1,103 to USD4,088 in 2005, to USD6,091 in 2012 [14]. The new emerging demand has added to continued growth from the US, Europe and other emerging countries. The global consumption of oil rose from 82.6 million barrels a day in 2004 to 85.6 million barrels a day in 2007 to 89.3 million barrels a day in 2012 [10]. Oil prices also depend on the US economy as the world's largest oil consumer. These factors contribute to a high price hike in July 2008 (Figure 1.2), with the US economy hitting a low. Figure 1.3 charts the annual US oil consumption against WTI price per barrel between 1973 and 2011.

[19] Per capita: per unit of population or per person.

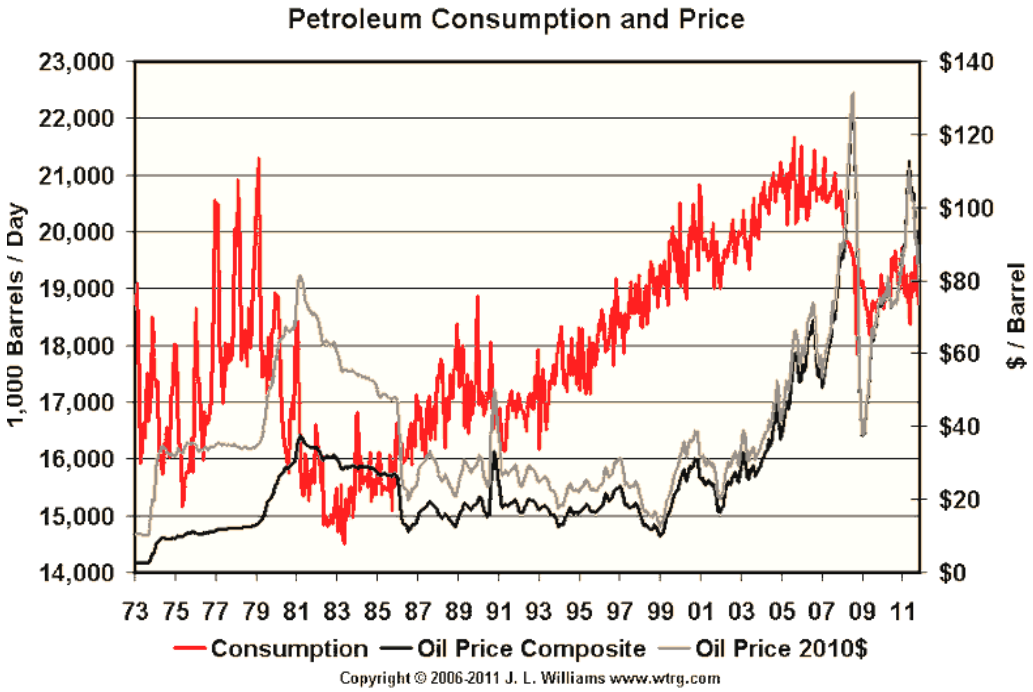


Figure 1.3 United States (US) Petroleum Consumption for Year 1973 to 2011 [9].

ii) Supply—The Organisation of Petroleum Exporting Countries (OPEC[20]) produces more than 30% of the world's supply. Any changes in OPEC's production policy contribute to a major effect on the crude oil price market. The years from 2005 to 2007 witnessed the reductions of OPEC oil production from 34.82 million barrels per day in 2005 to 34.18 million barrels per day in 2007. The OPEC oil production reductions, amongst other factors, caused drastic increases in oil prices in 2005 to 2007 hitting a peak price of USD147.27 on July 2008. OPEC gradually increased oil production in 2008 with an increase of 35.49 million barrels of production per day. Although OPEC plays a major role in determining the oil price in the market, there are 10 other oil-producing countries that supply more than 63% of production to the world [15]. Table 1.1 shows the production of these oil-producing countries with their share of world based on [16]. Russia, the US and China top the list by producing almost 40% of the world's production.

Table 1.1 The Top Non-OPEC Oil-producing Countries in the World (2013 est.).

NO.	OIL-PRODUCING COUNTRIES	PRODUCTION (Billion Barrel per Day)	SHARE OF WORLD (%)
1	Russia	10.9	13.28
2	United States	8.5	9.97
3	China	4.0	4.56
4	Canada	3.6	3.6
5	Mexico	2.9	3.56
6	Brazil	2.6	3.05
7	Norway	2.0	2.79
8	Kazakhstan	1.6	1.83
9	United Kingdom	1.1	1.78
10	Colombia	1.0	0.97

[20] OPEC: a permanent intergovernmental organisation of 12 oil-exporting developing nations that coordinates and unifies the petroleum policies of its member countries [116].

iii) World crisis—Crises in oil-producing countries affect oil prices by pushing the price up, and, given the fact that most oil-producing countries are prone to crisis, markets shift easily. When a crisis is anticipated, traders/suppliers will shift their productions down. This will trigger inflation[21] in the world economy which results in a higher oil price. Increases in oil price between November 2011 and January 2012 were due to Iran threatening to close the Straits of Hormuz after proof was discovered regarding their nuclear weapons program [17]. This would have disrupted the world's oil supply as the Straits are a major shipping route for exporting and importing. As Iran is one of the OPEC countries that contribute to more than 40% of production to the world, this threat significantly impacted investors which thus affected oil prices. Other world turmoil in 2011 also impacted the market when investors were concerned about the unrest in Arab countries. As the Middle East sits atop two-thirds of the world's reserves, this situation caused the price to increase to a peak of USD113 per barrel in April. Moreover, violence in Nigeria, Africa's top oil producer, had also caused the reduction of production by a quarter. The price went down to USD100 per barrel by June when the investors' concerns on supply disruptions disappeared.

iv) World Disasters—a dramatic natural or man-made disaster could also push oil prices up in the market. In 2005, Hurricane Katrina[22] and Hurricane Rita[23] caused major production disruptions with offshore oil and gas platforms destroyed and pipeline damages. US commercial crude oil inventories also decreased to 1.5 million barrels on August 2005 [18] since the natural disaster. Traders' concerns rose in May 2011 when Mississippi River floods threatened oil refineries. Their reactions were shown towards the market, pushing the price up in early-May. An explosion occurred in April 2010 on a BP-contracted oil drilling rig, the Deepwater Horizon in the Gulf of Mexico. The explosion, which raised environmental concerns regarding Gulf wildlife, caused prices to spike.

[21]Inflation: The rate at which the general level of prices for goods and services is rising, and, subsequently, purchasing power is falling [118].

[22] Hurricane Katrina: the deadliest and most destructive Atlantic tropical cyclone of the 2005 Atlantic hurricane season [122].

[23]Hurricane Rita: the fourth–most intense Atlantic hurricane ever recorded and the most intense tropical cyclone ever observed in the Gulf of Mexico [121].

The man-made disaster had also caused the price to increase from USD81.29 in March 2010, to USD84.58 in April 2010 as it caused damages to the oil refineries. These damages had caused disruptions to the level of inventories.

v) US Dollars (USD)—USD is used as a benchmark currency for marketing WTI price in NYMEX. For a commodity that trades globally, using a single-currency system reduces the cost of transactions. A declining USD means cheaper oil. It also increases demand from European and foreign countries, which leads to higher oil prices if the supply remains constant.

The crude oil market is full of the unexpected, and its prices respond swiftly to the events discussed in (i) to (v) of this subsection, with investors' anticipations playing large roles in determining the crude oil prices. Futures prices[29] reveal not only investors' expectations, but also their responses in current prices. Investors will take advantage of arbitrage[30] opportunities if current and futures prices are out of sync, by bringing prices back in line [19].

In this thesis, a study is conducted to derive better prediction tools that not only give insights into the behaviour of the market but also anticipate the sentiments or movements in the market in order to benefit researchers and investors in regulating prices. Ye, et al. in [20] suggested a simple model design that targets practitioners, wherein simplicity and ease of use will make a model attractive and useable for investigating different market scenarios. Section 1.3 will further discuss the significance of this research in order to develop prediction models.

[29] Future prices: agreement (contract) agreed between two parties to buy or sell crude oil of standardised quantity and quality for a price agreed upon today with delivery and payment occurring at a specified future date.

[30] Arbitrage: a trade that profits by exploiting price differences of identical or similar financial instruments. It exists as a result of market inefficiencies.

1.3 RESEARCH SIGNIFICANCE

Crude oil, as one of the most volatile commodities, is a real concern to the world in terms of its fluctuations. An attempt to understand (and model) its pattern of behaviour, though difficult, can lead to a possible prediction scheme. A constructive prediction study to anticipate crude oil price behaviour is essential to remedy such problems caused by geopolitics and environmental causes (i.e. natural disaster), as discussed in section 1.2 with subsections 1.2.1 to subsection 1.2.1.2.

1.3.1 The Domino Effects of Crude Oil Price on Economy

On a large scale, changes in oil prices have been associated with major developments in the world economy. Its high level of dependency on the factors discussed in subsection 1.2.1.2 is often seen as a trigger for inflation and recession. An increase in oil prices can cause an inflationary shock to an oil-importing country and oil production industries as its inward shift on aggregate supply gives pressure to the price level. A change in the crude oil price can also lead to income shifting between importing and exporting countries in terms of trade [21] as oil accounts for the most active trading commodity, with over 10% trade worldwide [22]. For net oil-importing countries, an increase in price will lead to higher inflation, downward pressures of exchange rates, increases in budget deficits, higher interest rates, and lower investment which will lead to a drop in national income. Vice versa, oil-exporting countries will gain more income if the price is increased because of its export gain. Nevertheless, a bigger price increase with sustaining higher price will demonstrate a bigger economic impact on the world's economy as this will ignite the domino effect in the commodity's market [23].

Increases in oil prices will generally increase the price of food, since food is dependent on transportation. The same goes for entertainment and other things. From the industry perspective, forecasting price movements is part of the decision–making process as it involves explorations into the development, valuation and production process influences. For a government, price predicting is actually an evaluation process of preparing export policy and maintaining national reserves [24]. Changes

in price will impact the government's revenue, especially if it is an oil-producing country.

The high dependency of crude oil price on other factors related to the market has caused market volatility and produced a major domino effect. Research with the objectives to systematically analyse the market's behaviour and to predict its future prices is essential as this will help practitioners calculate and minimise the risks of such impacts to an organisation or nation. As this research is aimed to benefit practitioners, this study aims to produce a practical and simple prediction tool that works as a 'decision support system' by imitating the way investors react to the market. Subsection 1.3.2 introduces the objectives of the research.

1.3.2 Research Objectives

The aim of this research is to develop a promising multi-variant mathematical model (in order to predict the crude oil prices) based on advances in machine learning and computational intelligence. Machine learning and computational intelligence approaches were used in this research to execute characteristics similar to a decision support system of a human, and subsequently to apply it to imitate human anticipation. The objectives of the research are:

i) data mining and extracting information via online news articles, in order to investigate and understand the market's behaviour. A knowledge-acquiring process identifies problems and processes them for solutions. It is an important phase as the knowledge acquired helps to construct a systematic method to support problem-solving;

ii) developing a quantitative prediction model by using numerical data, in order to evaluate the performance of the prediction based on the solutions provided by (i). This phase also helps to validate the extracted features retrieved from (i) which contain important events that regulate the crude oil price market;

iii) analysing the sentiments of news articles as a preliminary phase to process linguistic elements in a prediction model. The process imitates the way an investor anticipates events and reacts towards news that relate to the market. This phase is also an extension of the process done in (i) where it identifies events and maps the features that contribute to the problems in the market in a systematic form of algorithm or rule;

iv) building a knowledge and a rule base to store extracted linguistic information obtained from (iii) retrieved from the online news articles. The news articles containing events that affect the oil market are set as input;

v) building a dynamic expert system based on the rules obtained in (iii) which enables practitioners to easily improve the system as required and simply evaluate the market from time to time. The application of this dynamic expert system is aimed to be both practical and user-friendly;

vi) developing a prediction model with linguistic information stored in (iii) as input, in order to investigate the sensitivity of the model in reacting to the linguistic inputs obtained from the news. This imitates the reaction of investors in order to respond to ‘rumours’ received from news, and to investigate how these ‘rumours’ influence anticipations and expectations of current condition of the market; and, finally,

vii) integrating both (ii) and (vi) as a hybrid prediction model, in order to investigate the relevancy of linguistic data combined with quantitative data in deriving a better prediction model for the crude oil market. This phase imitates the way an investor calculates the risk of actions, based on the ‘facts’ (quantitative data) and ‘rumours’ (linguistic data) given in the market. This phase investigates how the combination of these two different forms of data helps to derive anticipations before reactions to the current condition of the market.

1.4 RESEARCH FOCUS

Although analysing the crude oil market is interesting, predicting its price is still a difficult task. Shin, et al. [23] agreed that predicting oil prices is a complex process due to oil's high dependence on other factors. Investigations were made into previous research about the importance of predicting the crude oil price accurately. It is also important, when developing prediction models, to take the practitioners' specifications into account as those individuals are involved in the market directly. A practical predictive modelling tool would help them in decision making. Although the crude oil price is popular as an everyday topic discussed everywhere, research in the area is still scarce. Hence, this thesis, with the objective to fill the knowledge gap, will focus on the predictive side of the domain with a preliminary analysis to develop a potential decision-making model (section 6.3).

The five main research areas focused on in this thesis are identified as:

1. The use of different factors in the market with WTI price as input. The price's volatility behaviour is presented by the dependency of the crude oil price market on other factors related, as discussed in subsection 1.2.1.1. Good interconnections and dependency of factors related to the crude oil price is hypothesised as providing high prediction accuracy. Neglecting these related factors discounts the prediction's credibility. The majority of authors in the literature review discussed in Chapter 2 have used WTI[33] and Brent[34] [25] crude oil prices as the only input for their prediction models. Other inputs that have also contributed to the market are absent. Thus, this thesis will utilise various factors discussed in subsection 1.2.1.1 as a reference to retrieve suitable factors for use as input in the prediction model.
2. The use of supply and demand factors as input. Other than crude oil prices, the next most common factors employed as input in the prediction model are demand and supply. Although demand and supply are main contributors to the market's volatility, the use of these observations alone is insufficient in propagating the information offered by the oil trends. In this thesis, supply

[33] West Texas Intermediate (WTI) price: traded based on high quality crude oil and a large contributor to refining large portion of petroleum.
[34] Brent price: traded based on a combination of crude oil from fifteen different oil fields located in the North Sea.

and demand will be included as among the important factors to be used in the prediction, together with other factors discussed in section 1.2.

3. Data pre-processing and data representation. Pre-processing and representing data in a normalised form were absent from previous research, as discussed in the literature review. Instead, data presented in time-series are still popular among the authors. Pre-processing and representing data into a normalised form in some degree will improve the prediction results, as the processes are meant to reduce errors and offer less noise, as well as decrease computational burden. In this thesis, all input data for the models will be pre-processed and represented in a normalised form, with analysis and evaluation of this usage conducted in the thesis.
4. Directional price vs. discrete price. Predicting the trend of crude oil prices is a popular topic in the literature. Thus, a prediction of both trend and discrete price will be more attractive for implementation in the real world. Based on both prices, this thesis will predict and evaluate the performance.

1.5 RESEARCH CONTRIBUTION

The thesis contributes to extant knowledge by developing five different models to predict the crude oil price market and adheres the objectives laid out in subsection 1.3.2. The five different models are:: i) the hierarchical conceptual (HC) model; ii) the quantitative prediction model; iii) the rule-based expert model; iv) the linguistic prediction model; and v) the linguistic-quantitative model. These models are discussed in-depth throughout the thesis with (i) and (ii) discussed in Chapter 3, (iii) and (iv) in Chapter 4 and (v) in Chapter 5. The contributions of these models to the thesis are summarised as follows:

1. The hierarchical conceptual (HC) model identifies a set of features that contribute to the volatility of the crude oil price. The performance of this set of data features was evaluated in the quantitative prediction model, based on the prediction's accuracy and minimal error production. Artificial Neural Networks Quantitative (ANN-Q) model is defined in the thesis as the quantitative prediction model, which validates the extracted features derived from the HC model, wherein the model produced a minimum prediction error of 0.00896 and 93.33% directional accuracy. Negnevitsky [26] suggests that a minimum error value obtained from a network of a prediction model is considered to have converged[37]. A sensitivity analysis was done to examine the relationship between different inputs (derived from HC model). The analysis proved that the minimum prediction error which interprets good interconnections exists between the extracted features used as inputs in the prediction network.
2. The process of grammar selection based on language structure, via Context Free Grammar (CFG), contributed to the building of a 'dictionary'. This domain-specific 'dictionary' has led to grammar derivation, enabling the annotation and extraction of grammar features of a *coreTerm*[38] and a *contentCategory*[39] in a fragment of an article.
3. Extracted grammar fragments (features) have contributed to new and important input data for creating a rule-based expert model and a linguistic

[37] Converged: interconnected inputs.
[38] *coreTerm*: variable used to annotate the features in a fragment.
[39] *contentCategory:* variable used to annotate the sentiment category i.e. *positive* or *negative* in a fragment.

prediction model in Chapters 4 and 5. The monthly aggregated sentiments derived from the extracted grammar fragments were contributed as the sole input in the linguistic prediction model. Meanwhile, the quantified sentiments obtained from the grammar fragments were contributed as input to delineate the rules into a decision tree. The aggregated and quantified sentiment derived from Chapter 4 are useful references by which to interpret the monthly events occurring in the market, which were then induced into a C4.5 algorithm to generate a decision tree as a set of rules.

4. The rules produced from the decision tree were converted into a fuzzy inference system as the set of fuzzy rules in order to establish the expert system systematically. The rule-based expert model, a fuzzy expert system, enables users to improvise the system by adding new knowledge (in terms of new rules) in the future.
5. News sentiments obtained from the extracted fragments in Chapter 4 were exploited as inputs in the linguistic model, which produced promising prediction results. This linguistic prediction model proves to have contributed to the anticipation process of a network through its capability of adding knowledge, even incomplete data. The exploitation of this knowledge gives insight into the reaction of the market and the impact on the price. The qualitative information extracted from the mined news articles complemented the quantitative element in the network, as it indicated events that happened quantitatively.
6. The linguistic information contained in the news sentiments complements the quantitative values contained in the quantitative model where it adds value into the hybrid prediction model even with small data. This validates the credibility of the domain-specific ‘dictionary’ in Chapter 4 as a good extractor in extracting the appropriate information, and has proved to contribute in deriving promising prediction results.

1.6 THESIS STRUCTURE

The thesis is composed by six main chapters summarised as follows:

CHAPTER 1: INTRODUCTION

This chapter will briefly discuss the research background, its significance and objectives.

CHAPTER 2: CRUDE OIL PRICE PREDICTION

Chapter 2 presents the approaches available in the literature to provide solutions to the problem of crude oil price prediction. The discussion begins with i) an overview of the crude oil price market, and, later, ii) a discussion of other approaches available for this crude oil price prediction.

CHAPTER 3: THE DEVELOPMENT OF A HIERARCHICAL CONCEPTUAL MODEL FOR ARTIFICIAL NEURAL NETWORK-QUANTITATIVE PREDICTION MODEL

This chapter discusses i) the development of the Hierarchical Conceptual (HC) model as the basis of information retrieval for the research, and ii) the development of the Artificial Neural Networks-Quantitative (ANN-Q) model to predict the price with quantitative data.

CHAPTER 4: CONTENT UTILISATION AND SENTIMENT ANALYSIS OF MINED GOOGLE NEWS IN FUZZY GRAMMAR FRAGMENTS WITH THE DEVELOPMENT OF FUZZY EXPERT SYSTEM FOR A RULE-BASED EXPERT MODEL

This chapter aims to extract the appropriate and relevant information from news articles to use as inputs in the linguistic model by i) applying the fuzzy grammar

fragment extraction, and ii) developing the fuzzy expert system based on the sentiments of the extracted textual information. The outcomes from this model will be implemented in the linguistic prediction model as the new input.

CHAPTER 5: PREDICTION WITH MINED SENTIMENTS FROM GOOGLE NEWS FOR THE LINGUISTIC MODEL AND ITS HYBRIDISATION WITH THE QUANTITATIVE PREDICTION MODEL

This chapter discusses i) the exploitation of sentiments as inputs for predicting the crude oil price, ii) the development of ANN modelling to predict with the linguistic information, iii) the hybridisation of both linguistic and quantitative data to predict, and, finally, iv) the discussion of empirical results gained from the models.

CHAPTER 6: CONCLUSION AND SUGGESTIONS FOR FUTURE RESEARCH

This chapter summarises the main achievement of this research. It identifies the main contributions of the models developed in this thesis. The chapter suggests some improvements that could be made to the existing models and research topics that might be interesting to cover in the future.

Chapter 2 Crude Oil Price Prediction

Overview

This chapter presents the approaches taken in the literature to provide solutions to the problem of crude oil price prediction. The discussion begins with an overview of the crude oil price market and, later, a discussion of other approaches to crude oil price prediction.

2.1 CRUDE OIL PRICE MARKET PREDICTION

Crude oil price market prediction is known for its obscurity and complexity. Due to its high uncertainty degree, irregular events, and the complex correlations with the market, it is indeed difficult to predict its movement. Panas and Ninni [27] mentioned that the crude oil market shows strong evidence of chaos and has developed as one of the most volatile markets in the world. Despite its chaotic behaviour, though, researchers have found this to be an interesting area to explore. Crude oil price market predictions are mainly derived from such approaches as i) single statistical and econometric models; ii) a single Artificial Intelligence (AI) model, and iii) the hybrid approaches.

2.1.1 Single Statistical and Econometric Prediction Model

Statistics and econometric approaches are popular in predicting crude oil market prices.

The Generalized Autoregressive Conditional Heteroskedasticity **(**GARCH) model [28] and the Naive Random Walk [29] [30] are among the statistical and econometric models used to predict crude oil prices. Abramson and Finniza [31] successfully utilised a probabilistic model to predict oil price. The research was conducted based on a case study of the probabilistic inheritance of Belief Network (BN) models. The models are used to forecast crude oil price and then produce a probabilistic prediction [32]. The probabilistic prediction is actually generated by running a Monte Carlo[43] analysis on annual West Texas Intermediate (WTI) average prices.

For the purpose of the simulation experiment in [32], the analysis done in this study is based on two assumptions: i) the timing of Iraq's return to the market; and ii) the impact of oil exports from the former Soviet Union. Subsequently, three variables from the input vector are then used to define the events: i) the probabilities of embargo ends; ii) total demand; and iii) other world productions.

Their research was conducted based on a case study in the use of probabilistic Belief Network (BN) models or Influence Diagrams (ID), introduced as graphical mechanisms for automated Bayesian inference and expected utility calculations. Their BN rules showed that a mathematical approach with respect to existing belief can be modified to take new evidence into account. Their probabilistic prediction was generated by running a Monte Carlo analysis on annual WTI average prices for years 1993, 1994 and 1995. This research developed two projects: i)ARCO[44]1, a knowledge-based system that uses a BN model to generate quarterly average price; and ii) ARCO2, a BN-based system developed for intermediate and long-term

[43] Monte Carlo analysis: computerised and mathematical technique that allows people to account for risk in quantitative analysis and decision making.

[44] ARCO: Atlantic Richfield Company

forecasting. This project is motivated to integrate political and economic variables within a single model. This paper focused on the development of ARCO2.

In ARCO2, three input nodes were selected to define a set of scenarios on the simulation experiment:

i) the end of the embargo (with the probability of Iraq rejoining the market in 1993, 1994 and 1995);

ii) the total demand (with the probability distribution describing demand); and

iii) other world production (with the probability distribution describing production outside the six core oil-producer countries and the former Soviet Union).

Two distinct probability distributions were defined for the embargo end, and three each for total demand and other world production, tabulated in Table 2.1:

Table 2.1 Probability Distribution for Three Input Nodes to Define Probability Sets in ARCO2

Embargo	**Demand**	**Production**
2	3	3

Eighteen identical distinct forecasting networks were defined for the various combinations of distributions. Each network forecasted based on a Monte Carlo analysis over three years of annual averages of 1993, 1994 and 1995 was instantiated by 1000 instances of the network. Hence, 54,000 possible future scenarios for the market were analysed. Their research [32] showed the annual average prices represented by the quantiles[47] are almost certain to remain good predictive patterns. Trends are indicated by their forecast; these trends are substantially more important than the forecast itself. The result from their simulation is robust and consistent with the market's annual average prices, between USD15 to USD25 per barrel. Out of 54,000 events applied in the model, only 0.75% incurred an error. In conclusion, the most significant result of their entire simulation was that all 17 other networks predicted the same price patterns as described in ARCO1.

[47] Quantiles: values which divide the distribution such that there are a given proportion of observations.

Another statistical model prediction made for crude oil price was made by Morana [33]. This research used a semi-parametric approach suggested in [34] for short-term oil price predictions. It also used GARCH[49] to employ oil price changes to predict the oil price distribution over a short-term horizon. The approach used one-month-ahead daily Brent oil price which spanned periods with high uncertainty (November 21, 1998 to January 21, 1999). Furthermore, the forecasting analysis is based on the last two months of the available data, and, according to the analysis, the results were different from the actual. This is may be linked to the widening of the forecast confidence interval. Nevertheless, the study offers improvement from [34]. The main features of these data selection are, first, it does not display a global trend, and, second, it displays alternating periods of high volatility followed by periods characterised by relative tranquillity (non-volatility). The oil price is first modelled using the Martingale process[50] [35] and later compared with the Random Walk[51] model. The Martingale process is interpreted by the following equation:

$$y_t = y_{t-1} + \varepsilon_t, \tag{2.1}$$

Where ε_t is a mean-zero that a linear-independent distributed type of constant random variable while, y is the oil price in period t. This error term features its difference from the Random Walk model. However, the performance of both models is poor in terms of bias and fails to predict the signs of oil changes in more than 50% of the cases. The methodology first analyses the in-sample parametric (variables) and then analyses the out-of-sample non-parametric. Empirical distribution functions are obtained by repeating the procedures that can be used to compute the quantiles of interest. Eighty percent of forecast confidence intervals are selected to compute the quantiles for 48 one-month forward price predictions. The results which were carried out using the Brent oil price series show a deteriorating prediction ability and fails to satisfactorily track the oil price trend. This failure is related to the widening of its confidence interval. It is suggested that this method be used to obtain the expected performance measure of the forward price as the predictor for the oil price. Finally,

[49] GARCH: Generalized Autoregressive Conditional Heteroskedasticity.
[50] Martingle process: a stochastic process whose expected value at each step equals its previous realisation or observed value.
[51] Random Walk: a mathematical formalisation of a path that consists of a succession of random steps.

this approach suggests an intuitive result which enables the quantification of the volatility surrounding the oil price prediction of the methodology used.

Another statistical model used in the research is by Ye, et al. [20] in predicting a monthly WTI spot price[55] using relative inventories. This research used the Relative Stock[56] (RSTK) model as the basis to predict the price by comparing two other alternatives models: the Naïve Autoregressive (NAIV) forecast model and the Modified Alternative (MALT) model. This research focuses on the concepts of normal and relative levels of petroleum market variables which are similar to the national rate theory implemented in macroeconomics modelling and empirical studies. The normal inventory level is composed by normal demand and normal operating requirements, whereas the relative inventory level is the response to market fluctuations. The normal inventory level is calculated by de-seasonalising and de-trending the historical data, while the relative inventory level is defined by the actual inventories deviation determined from the historical normal level. Concurrently, the relative inventory level (RIN) is denoted by the equation (2.2):

$$RIN_t = IN_t - IN_t * \tag{2.2}$$

where IN_t is the actual industrial OECD[57] petroleum inventory level in month *t*, and IN_t* is the normal level. Letting D_k, k = 2, 3, .. 12 be 11 seasonal variables and T be the linear trend, the normal inventory level is calculated equation (2.3):

$$IN_t * = a_0 + b_1 T + \sum_{k=2}^{12} b_k D_k \tag{2.3}$$

where a_0, b_1, and b_k, k = 2, .., 12 are estimated coefficients from the de-trending and de-seasonalising the observed total petroleum inventory. The study limits its prediction period from January 1992 to April 2003. This limitation is implemented to avoid the impact the first Gulf War had on markets and to limit the analysis to a

[55] Spot price: The current price at which a particular security can be bought or sold at a specified time and place.
[56] Relative stock model: A momentum investing technique that compares the performance of actual stock from a historical level to that of the overall market.
[57] OECD: Organisation for Economic Co-operation Development.

consistent data series. Before the prediction process started, a number of factors were considered by [20] in order to prepare for the best prediction model. Specific criteria were laid out by [20] to maximise the model's performance:

i. WTI price was found to have a unit root that diminished the predicting ability. The WTI price and inventory are not used in the analysis;
ii. Crude oil inventory and oil product inventories are found to perform better than crude oil inventory alone.
iii. Series of low stock indicators are not included in the analysis to capture the asymmetric market behaviour. The absence of low stock indicators in the analysis is found to marginally reduce the in-sample prediction error and increase out-of-sample prediction error.
iv. Three lag lengths for the relative inventory variables plus the current period of relative inventories are investigated and included. The three lags included are R^2 [58], AIC[59] and SBC[60] as well as the in-sample and out-of-sample prediction results.
v. GDP potential role is not included in the investigation as it is found to have no source of future prediction for a monthly OECD GDP aggregate. The inclusion of US GDP does not sufficiently provide good prediction results in the model.

Therefore, the best specification to be used in the research is the RSTK model to short-run predicts the monthly WTI spot price with WTI price as input. The RSTK model is formulated as follows:

$$WTI_t = a + \sum_{i=0}^{3} b_i RIN_{t-i} + \sum_{j=0}^{5} c_j D_j 911 + dLAPR99 + eWTI_{t-1} + \varepsilon_t \qquad (2.4)$$

with *t* for the *t*th month and *i* for the *i*th month prior to the *t*th month. a, b_i , c_j, d and e, $i=0, 1, 2, 3$ and $j=0, 1, 2, .., 5$ referring to the 6 months from October 2001 to March 2002, the coefficients to be estimated. While D_j911 is a set of single monthly

[58] R^2: coefficient of determination of variability in data set.
[59] Akaike Information Criterion (AIC): a measure of the relative quality of a statistical model for a given set of data.
[60] Schwarz's Criterion: An index used as an aid in choosing between competing models.

variables to account for market disequilibrium[61] following the September 11, 2001 terrorist attacks in the United States, the LAPR99 is a level shifting variable corresponding to the effect that OPEC quota tightening had on the petroleum market in the beginning of April 1999. The in-sample prediction used a root mean squared error (RMSE), mean absolute error (MAE) deviation, mean absolute percent error (MAPE) statistics and the Theil *U* inequality coefficient. Moreover, the bias, variance[62] and covariance[63] proportions are used for the three models in the estimation sample period. From the prediction experiments, RSTK gives the best prediction result with 1.538 and the smallest value of root mean squared error (RMSE) compared to the other two alternative models. The comparison of the dynamic forecasting can be seen in Figure 2.1.

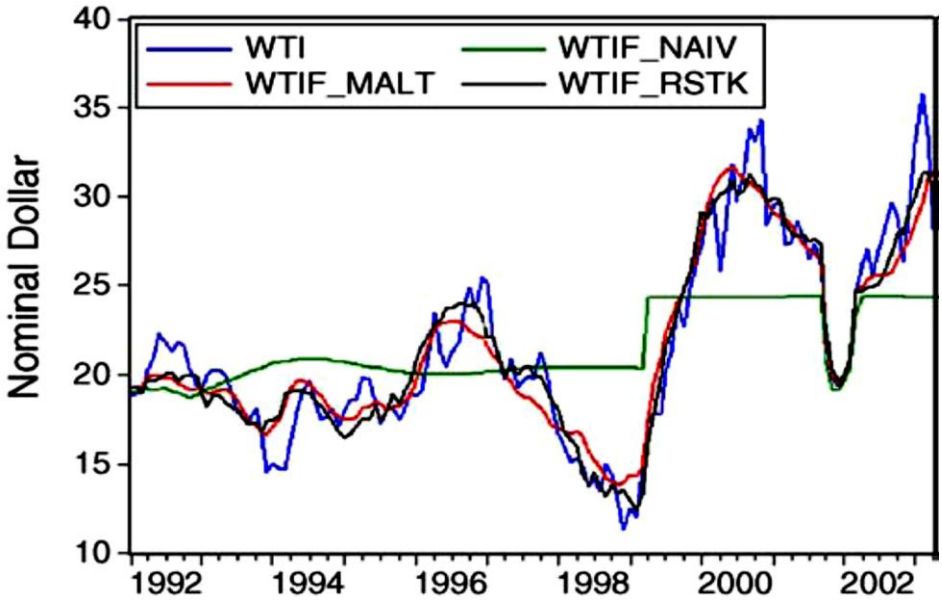


Figure 2.1 The Comparison of Dynamic Forecasts and West Texas Intermediate (WTI) Price [20].

Concurrently, the prediction made by the mentioned models in this section 2.1.1 show good prediction results in studies using linear or near linear data.

[61] Disequilibrium: Loss or lack of stability or equilibrium.
[62] Variance: A measurement of the spread between numbers in a data set. The variance measures how far each number in the set is from the mean.
[63] Covariance: A measure of the degree to which returns on two risky assets move in together. A positive covariance means that asset returns move together. A negative covariance means returns move inversely.

Conversely, crude oil prices are literally related to nonlinear and irregular events. Some experiments demonstrated that the predictions made in the statistical models perform poor results [33] if continuing as statistical and econometric models meant to handle linear assumptions and not the nonlinear patterns. Therefore, nonlinear and artificial intelligence (AI) models are introduced in section 2.1.2 to overcome these limitations.

2.1.2 Single Artificial Intelligence (AI) Prediction Model

Nowadays, AI models are among the popular tools used for prediction. As an alternative tool to statistical and econometric models, AI offers recognition ability for complex patterns, and it also provides intelligent reasoning and intelligent decision-making based on data. Support Vector Machine (SVM) was used as an alternative tool to the other single prediction models, to predict crude oil prices. One prediction models that used an AI approach for predicting the crude oil price is by Xie, et al. [36]; where, for the task of time-series prediction, this research focused only on the Support Vector Regression (SVR) model. To evaluate the prediction ability of SVM, it was compared with Autoregressive Integrated Moving Average (ARIMA) model and Back Propagation Neural Network (BPNN). For the task of time-series prediction, this research [36] focused only on Support Vector Regression (SVR). In SVM, a linear function is always used to solve the regression problem, but, when dealing with nonlinear regression, SVM maps the data x into a high-dimensional feature space via a nonlinear mapping φ and makes a linear regression in this space. The equations used in the study are based on Vapnik's ε-insensitive loss function[73] for the goal of regression and time series prediction. It is formulated as below:

$$\psi(f(x)-y)=\begin{cases}|f(x)-y|-\varepsilon & \text{for } |f(x)-y|\geq \varepsilon \\ 0 & \text{otherwise}\end{cases} \tag{2.5}$$

where $\psi(.)$ denotes a cost function, $f(x_i)$ the linear function and ε denotes the loss function. The procedure of SVM for this time series prediction is illustrated as Figure 2.2

[73] ε-insensitive loss function: method to optimise the generalisation bounds a loss function that ignores errors.

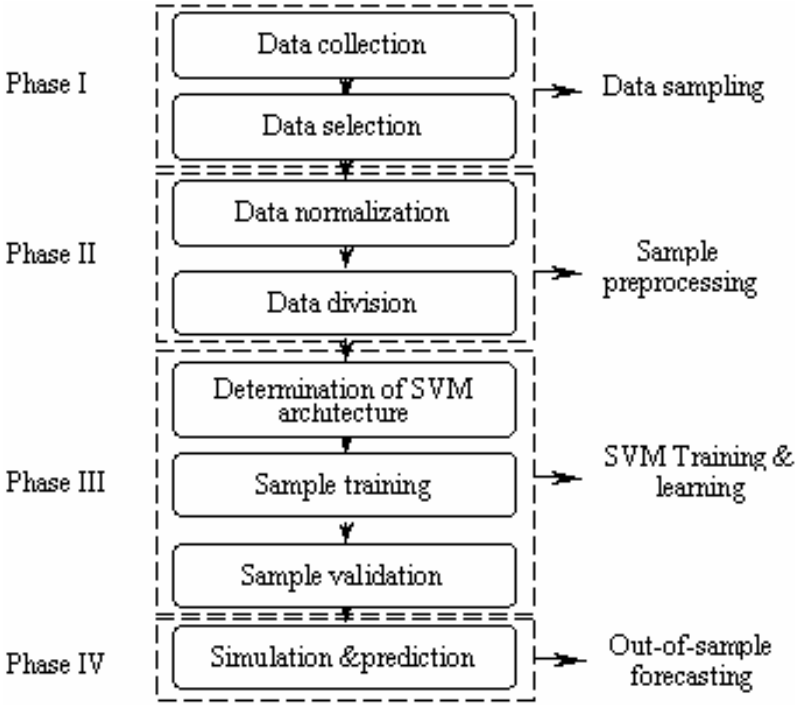


Figure 2.2 Support Vector Machine (SVM)-based Forecasting System Procedures [36].

Referring to Figure 2.2, the procedures of developing the prediction model involves:

i. Data sampling. Data used in this study are monthly spot prices of WTI crude oil ranging from January 1970 to December 2003 with a total of n = 408 observations;

ii. Sample pre-processing. No normalisation is used for this study for simplicity as SVM are resistant to noises due to the use of a non-linear kernel and ε-insensitive band;

iii. SVM training and learning vectors are the monthly data from January 1970 to December 1999 used for in-sample data sets (including 60 validation of data) with 360 observations for training and validation purposes; and

iv. Out-of-sample forecasting: 48 observations from the out-sample data for testing purposes.

The result of the experiment is presented in Table 2.2:

Table 2.2 Crude Oil Price Forecasting Results based on Support Vector Machine (SVM) [36].

Methods	Criteria	Full period	Sub-period I (2000)	Sub-period II (2001)	Sub-period III (2002)	Sub-period IV (2003)
ARIMA	*RMSE*	2.3392	3.0032	1.7495	1.9037	2.4868
	D_{stat} (%)	54.17	41.67	50.00	58.33	66.67
BPNN	*RMSE*	2.2746	2.9108	1.8253	1.8534	2.3843
	D_{stat} (%)	62.50	50.00	58.33	66.67	75.00
SVM	*RMSE*	2.1921	2.6490	1.8458	1.8210	2.3411
	D_{stat} (%)	70.83	83.33	50.00	58.33	91.67

The result from the research indicates that the SVM outperforms the other two models in terms of both RMSE and D_{stat} and this implies a good prediction tool for crude oil prediction. This research states that the D_{stat} indicator is a more important parameter as it is a more practical measurement than RMSE from the practical application point of view. This can reflect the movement trends of oil prices and will help traders to hedge their risk to make good trading decisions in advance. To conclude, based on the experiment, the SVM approach performs best with test results of 2.1921 and 70.83% for RMSE and D_{stat} respectively.

Among other single AI prediction models is the prediction developed by Abdullah and Zeng [37]. The model utilises a hierarchical conceptual (HC) model and an artificial neural network quantitative (ANN-Q) model via a back-propagation neural networks (BPNN) tool and has successfully validated the effectiveness of data selection process by HC[75]. A systematic approach based on online news mining was developed to retrieve information relating to the events involved in the crude oil price market. To understand the uncertainties involved in the market, factors (or features) that contributed to these uncertainties are first retrieved from the online news. The ANN-Q model successfully extracted a comprehensive list of key factors that caused the volatility, consisting of i) demand; ii) supply iii) economy; iv)

[75] Hierarchical Conceptual (HC): a set of concepts arranged in a tree structure diagram associating each concept with instances of that concept.

inventory; and v) population as the main features of the model. Each key factor mentioned in (i) to (v) contained subordinates. The sensitivity analysis made on the extracted features produced good interconnections of input portrayed by the promising results evaluated by the minimum normalised mean squared error that it produced. Hu, et al. [38], through their research on applying ANN in oil futures, agreed on the appropriateness of ANN in forecasting crude oil future prices. With a comparison made through 3 different models—Elman recurrent neural network (ERNN), recurrent fuzzy neural network (RFNN) and multilayer perceptron (MLP)—it was found that RFNN had the best predictive power.

2.1.3 Hybrid Prediction Model

Hybrid models were introduced by previous studies as remedies to the research problems encountered by single prediction models. This section discusses the numbers of studies that induced the hybridisation method of both statistics and artificial intelligence techniques into a hybrid model, as an alternative solution to the problems encountered in past research.

Wang, et al. [39] introduced TEI@I methodology in order to hybridise four models, combining:

i) Text mining (T);

ii) Econometrics (E);

iii) Intelligence (intelligent algorithm-I) components; and,

iv) Integrating (@[77]) of the mentioned in (i), (ii), (iii); and (iv).

This study integrates four different models: i) web-based text mining (WTM); ii) auto-regressive integrated moving average (ARIMA); iii) artificial neural networks (ANN); and iv) rule-based expert system (RES), to predict the price. ARIMA and ANN were used to model the linear and non-linear components of crude oil price time-series respectively. Concurrently, the irregular and infrequent events on crude oil price were explored by WTM and RES techniques. The methodology framework of this method is presented in Figure 2.3.

[77] "@": the symbol is introduced to emphasise the central role and function of integration.

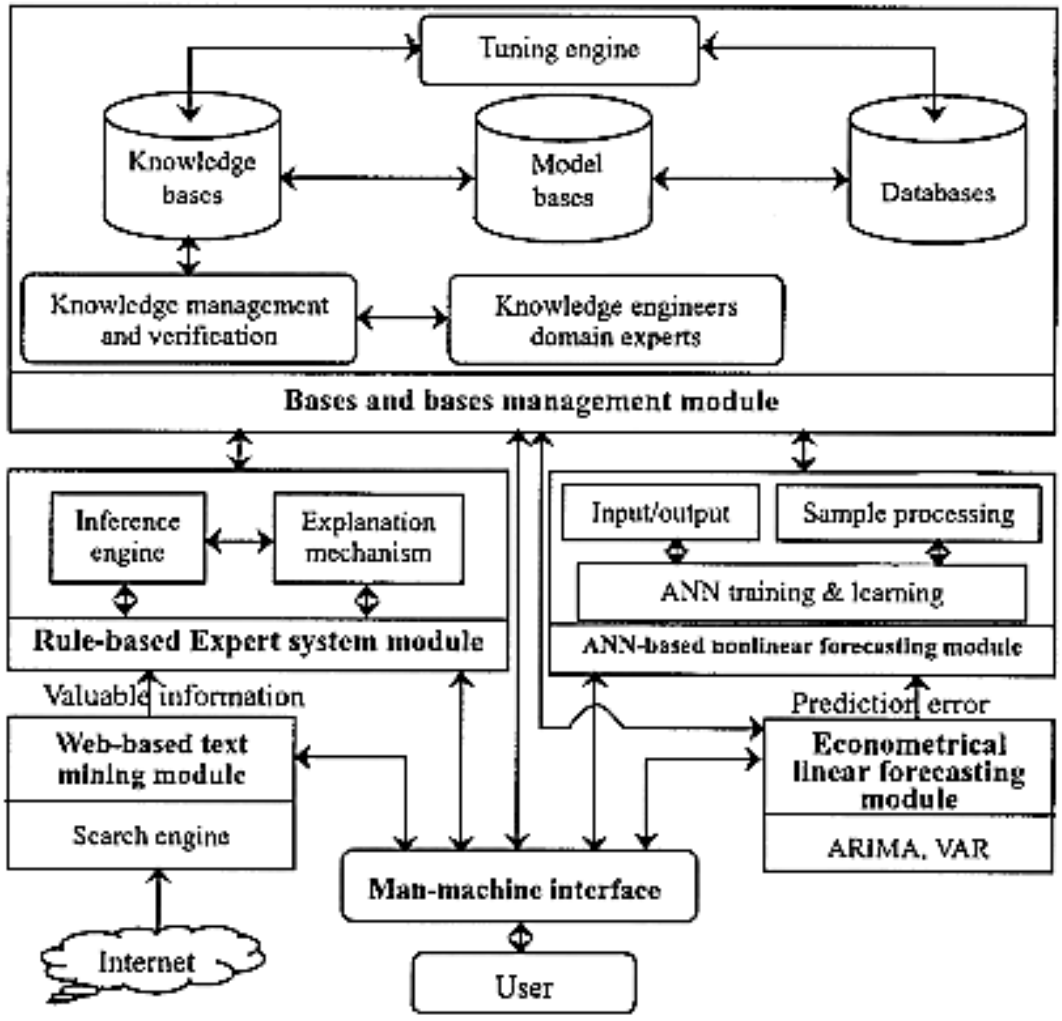


Figure 2.3 The TEI@I Methodology Framework for Crude Oil Price Forecasting [39].

The integration modules presented in Figure 2.3 is described as follows:

i) The man-machine interface (MMI) module—a graphical window acts as a hub of communication between users and the system, by exchanging inputs and the output data;

ii) The web-based text mining (WTM) module—it is the major component in the methodology. It collects information affecting oil price variability from the Internet and provides a Rule-based expert system (RES) with useful information to analyse the effects on the oil price;

iii) Rule-based expert system (RES) module—it constructs a knowledge base (KB) that represents all rules collected from knowledge engineers and domain experts. It is also required to extract rules to judge abnormal variability by concluding the relationships between volatility and the irregular key factors affecting it;

iv) Artificial neural network (ANN) module uses back propagation neural network (BPNN) as the learning algorithm which was developed based on [40]. This model is expected to capture the nonlinear characteristics of the time-series; and

v) RES module, ANN-based nonlinear forecasting module and MMI module, which use the Bases and Bases Management (BBM) module to fine-tune the knowledge gained from the ANN forecasting result, in order to adapt to a dynamic situation. The BBM module is important for this TEI@I approach as it provides strong connections between other modules in the approach, by aggregating algorithms, data and models from other modules.

To improve the forecasting performance, a novel nonlinear integrated forecasting approach is proposed by [39]. Firstly, ARIMA is used to fit the linear component of the time series by using the following equation:

$$e_t = Y_t - \hat{E}_t \tag{2.6}$$

where, Y_t is the actual value of time-series which assumed to be $\{Y_t,\ t = 1, 2, .., n\}$ and $\hat{E}_t$, as the forecast value of the linear components. Thus, the series of the

nonlinear components is represented by e_t. Next, the BPNN is trained by the previously generated nonlinear time-series components, and it can be seen as an error-correction process of the previous ARIMA module which is formed by the following mapping function. This will then generate I_t,, the forecast value of nonlinear components.

$$e_t = f\left(e_{t-1,} e_{t-2,}.., e_{t-p}\right) \tag{2.7}$$

In the third phase, the value of $\hat{T}_t$ is defined by the irregular effects of expected future events on oil price that are obtained from WTM and RES modules. Then, the three components of $\hat{E}_t, \hat{I}_t$ and $\hat{T}_t$ are combined in equation (2.8) to obtain a synergetic effect in time-series forecasting.

$$\hat{Y}_t = f\left(\hat{T}_t, \hat{E}_t, \hat{I}_t\right) \tag{2.8}$$

To verify the effectiveness of this hybrid model, a simulation was undertaken by using monthly West Texas Intermediate (WTI) spot price, ranging from January 1970 to December 2003 (408 observations) as the crude oil price data. From the total data, about 88% were used for data training and validation with the remainder of the total for data testing purposes. Later, the simulation conducted had produced a root mean squared error (RMSE) value of 1.0549. The same simulation also produced D_{stat} value of 95.83 % accuracy for the full period of 2000 to 2003. Simultaneously, some irregular events were found from WTM module via the Internet. The irregular events which were described as patterns in the research [39] contributed to the classification of events (or rules) into pattern groups, according to the specific period in the events. The group were divided into i) military, economic and political factors; ii) OPEC policy; iii) non-OPEC policy; iv) natural disaster; v) world economy; and, vi) other factors. In these six groups, there are 25 patterns reflecting the irregular events in the market. Wang, et al. [39] shared the same objectives as this thesis wherein different models were used to derive a more practical prediction tool as compared to models in [32] to [38]. Wang, et al. [39] also suggested that the change in trend is more important than a precision level of an outcome. Therefore, a directional change statistic was introduced to compute this. The proposed nonlinear integration forecasting approach introduced in this research [39] performed well in

predicting the price and is competitive with [12] in terms of the directional accuracy measured by the D_{stat} value. This is due to the integration of different models that generalise the different aspects of data used for the input. The evaluations made between the single model and the integration model also proved the success of this hybrid model. The integration of these modules generates synergy that made the research more accurate, which was absent in the single models. Furthermore, based on the result, the research suggested a profit opportunity in that using this hybrid model to predict the crude oil price in practice is a good practical indicator to be implemented by investors.

The Empirical Mode Decomposition (EMD)-based Neural Network (NN) ensemble learning paradigm for crude oil prediction by Yu, et al. [41] also suggests a promising AI method for predicting the crude oil price. The research model was evaluated using evaluation metrics: RMSE and the D_{stat} which both showed the outperformance of this model against other alternative models. Two main crude oil price series, WTI crude oil spot price and Brent crude oil spot price, were chosen as experimental samples for this research. The process of developing this technique can be divided into three steps:

i) first, the original crude oil price series were decomposed into a finite, and often small, number of intrinsic mode functions (IMFs);
ii) second, a three-layer of feed-forward neural network (FNN) is used to model each of the extracted IMFs. The model is used to ensure the tendencies of the IMFs in predicting accurately; and
iii) finally, the prediction results of all IMFs are combined with an Adaptive Linear Neural Network (ALNN), to formulate an ensemble output for the original crude oil price series.

In EMD, the processes involved are as follows:

i) for all local extrema, including local maxima and local minima, the data series of *x(t)*, are identified;

ii) all local extrema were regressed by a cubic spine line to generate its upper and lower envelopes of $x_{up}(t)$ and $x_{low}(t)$;

iii) the point-to-point envelope mean *m(t)* from upper and lower envelopes $(x_{up}(t) + x_{low}(t)/2)$ are computed;

iv) the details, $c(t)=x(t)-m(t)$ is extracted; and,

v) the properties of *c(t): i* if *c(t): (i) if c(t)* meets the two above requirements, an IMF is derived and meantime replace *x(t)* with the residual $r(t)=x(t) - c(t)$ are checked. If *c(t)* is not an IMF, replace *x(t)* with *c(t).*

The process is repeated until the stop criteria are satisfied, and, at the end of the process, the data series *x(t)* are expressed by:

$$x(t) = \sum_{j=1}^{n} c_j(t) + r_n(t) \tag{2.9}$$

Where *n* is the number of IMFs, $r_n(t)$ is the final residue and $c_j(t)(j = 1, 2, .., n)$ are the IMFs that have nearly zero means. Meanwhile, to model the decomposed IMFs and the residual component, a standard three-layer FNN trained with the error Back Propagation algorithm is selected, and, to aggregate the results produced by the FNN, ALNN[79] is chosen based on individual IMFs. The final output of the FNN-based forecasting model and the ALNN a single layer NN, that is used to combine the prediction result from FNN can be represented by equations (2.9) and (2.10) respectively:

$$f(x) = a_0 + \sum_{j=1}^{q} w_j u(a_j) + \sum_{i=1}^{p} w_{ij} x_i \tag{2.10}$$

$$f(x) = \varphi\left(\sum_{i=1}^{m} w_i x_i + b\right) \tag{2.11}$$

where x_i $(i=1, 2,.., m)$ represents the input variables, *f(x)* is the output, *b* the bias, w_i $(i=1, 2, .., m)$ the connection weight, *m* is the number of input nodes, and $\varphi(.)$ is the

[79] Adaptive Linear Neural Networks (ALNN): a simple two-layer neural network with only input and output layers and a single output neuron.

transfer function of the single-layer ALNN. The data used in this study were divided into WTI crude oil price and Brent crude oil data with daily data from January 1st 1986 to December 31st, 2000 used for training and the remainder for testing. Explicitly, data from May 20th, 1987 to December, 31st 2002 were used for training and data from January 1st, 2003 to September 30th, 2006 as the testing set. The prediction results are evaluated by root mean squared error (RMSE) and presented in Table 2.3. This research used the EMD-FNN-ALNN methodology, and has the best RMSE value for both WTI and Brent crude oil price compared to other methodologies. The errors produced are satisfactory with only one RMSE valued more than 2.000. In a comparison of methods (1) to (4) in Table 2.3, ALNN model improves WTI price results better when integrating EMD with FNN, and when integrating EMD with ARIMA as compared to integrating them with the Averaging model. Similar results were obtained when predicting with the Brent price. The single models introduced in the research also showed that the absence of synergy in the single model impacts their ability to predict accurately on both prices.

Table 2.3 The Root Mean Square Error (RMSE) Result from Simulation Experiments.

NO.	METHODOLOGY	WTI		BRENT	
		RMSE	RANK	RMSE	RANK
1	EMD-FNN-ALNN	0.273	1	0.225	1
2	EMD-FNN-Averaging	0.509	2	0.457	2
3	EMD-ARIMA-ALNN	0.975	4	0.872	4
4	EMD-ARIMA-Averaging	1.769	5	1.392	5
5	Single FNN	0.841	3	0.743	3
6	Single ARIMA	2.035	6	1.768	6

Another hybrid model relating to crude oil price forecasting is a rough-set refined text mining approach by another work of Yu, et al. [42], where text mining and rough-sets are combined to produce useful knowledge that can be used to configure and predict the tendency of the crude oil market. The advantage of this approach is that it can consider both the quantitative and the qualitative factors. The model input variables are all possible events that affect the crude oil market. The events were

extracted from the Internet and the internal file system, using the rough-set refined text mining approach. Other than that, world oil demand and supply, crude oil production and crude oil stock level were selected as input variables with monthly WTI price as the output variable. This approach has developed a promising tool for predicting the movement of crude oil market where it outperformed other models in its evaluation process.

Finally, research in [43] integrates empirical mode decomposition (EMD) with feed-forward neural network (FNN) and adaptive linear neural network (ALNN), EMD-FNN-ALNN, to formulate an ensemble output for the original crude oil price series. This study used daily WTI and Brent oil price ranging from January, 1986 to September, 2003, excluding public holidays. From the experiment, an evaluation was made, and Wang, et al. [43] concluded that this method offers an alternative prediction tool to crude oil price forecasting. Wang, et al. [43] also proved that the decomposition and ensemble techniques used in EMD (Decomposition)-FNN (Prediction)-ALNN (Ensemble) had improved the limitations of other previous single models. Details of this approach can be found in [43].

2.2 CONCLUSION

Given the literature reviewed so far, a number of problems can be identified. First, the data used in the predictions are typically drawn from the WTI price or Brent price, and they do not take into consideration other inputs aligned to the market. The volatility of the crude oil price market is a result of the dependency of the market on various factors. Neglecting these factors in predicting the market can downgrade the credibility of a prediction tool, preventing it from being comprehensive. A model with good prediction results demonstrates good interconnections between inputs and the output which suggests the state of dependence.

Secondly, studies that emphasise the volatility aspect of the market are still limited. The majority have focused on the price side of the prediction rather than the factors that caused the movements. Among other popular impact factors used in the crude oil prediction models are demand and supply. Although oil demand and supply play vital roles in the volatility of the price, the use of these observations only limits the potential of other factors such as input data, resulting in a model not being comprehensive. By including and connecting the key factors involved, a more comprehensive prediction of the market can be achieved.

Third, most of the research studied have utilised time-series data. Data pre-processing[81] and data representation[82] process were absent in most of the research. These two processes help to clean and reduce noises in data sets and normalise them in condensing the process of prediction, and, later, these help to generate accurate results. Without these processes, the prediction tool will be less reliable.

Fourth, studies have shown that predicting the prices' trends is more popular than predicting the discrete price itself. Discrete price predictions will make research more attractive and practical for practitioners even though the practicality of the studies conducted, to date, is still questionable. Table 2.4 summarises the models discussed in this chapter with the data used for the predictions.

[81] Data pre-processing: cleaning of data from missing values and noises.
[82] Data representation: conversion of data into a form that can be processed by a computer.

Table 2.4 The Impact Factors used by Different Models for Crude Oil Price Prediction Model.

NO.	MODEL	METHODOLOGY		INPUT (IMPACT FACTORS)	OUTPUT (PRICE)
1	Belief network [31]	Single	Statistics/ econometrics	politics, economics	WTI
2	Probability distribution [32]	Single	Statistics/ econometrics	OPEC policy, US demand, world production	WTI
3	Semi-parametric-GARCH approach [33]	Single	Statistics/ econometrics	Brent (price)	Brent
4	Relative Inventories [20]	Single	Statistics/ econometrics	crude oil inventory, oil product inventory	WTI
5	SVM [44]	Single	AI	WTI	WTI
6	ANN-Quantitative with BPNN [37]	Single	AI	Productions, proved reserves, no. of well drilled, consumptions, stocks, imports, forex, GDP[85], inflation, CPI[86], and population	WTI
7	RFNN [38]	Single	AI	WTI	WTI
8	TEI@I [39]	Hybrid	AI, statistics, econometrics	WTI	WTI
9	FNN-ARIMA-ALNN [41]	Hybrid	AI, econometrics	WTI + Brent	WTI + Brent
10	Rough-set refined text mining [42]	Hybrid	Data mining, statistics	Any possible events related to market (linguistic), demand, supply, production, stock	WTI
11	EMD-FNN-ALNN [43]	Hybrid	AI	WTI + Brent	WTI + Brent

[85] GDP: Gross Domestic Products.
[86] CPI: consumer price index.

Chapter 3 The Development of a Hierarchical Conceptual Model for the Artificial Neural Network-Quantitative Prediction Model

Overview

The crude oil market is known for its uncertainty. Its volatility creates mixed sentiments and draws interest from both researchers and practitioners. Previous statistical and econometric techniques used for prediction offer good results when dealing with only linear data. Nonetheless, the oil price market has large amounts of nonlinear data produced by irregular events. The continuous use of these techniques for prediction might demonstrate reduction in prediction performance where a dynamic model for predicting would be more appealing. In this chapter, a price prediction model based on an Artificial Neural Network (ANN) application is built to fit this purpose with the exploitation of quantitative elements as the main input. This chapter will discuss the development of a Hierarchical Conceptual (HC) model in exhibiting the basis of the research, and the development of Artificial Neural Networks-Quantitative (ANN-Q) model in applying the knowledge that ANN offers for domain prediction.

3.1 INTRODUCTION

Oil is one of the most fragile and volatile commodities in the world. Its volatility caused a sudden economic crisis through its drastic increase in July 2008 [45], resulting in a world recession. Starting with USD69 per barrel in April 2006 and later rising to US134 per barrel in July 2008, this phenomenon had an impact on oil-importing and oil-exporting countries. The crude oil market strongly affects world markets in terms of economics, politics and society. Contributing to over 50% of petroleum price, the market's volatility has been seen as impacting the world economy since petroleum is one of the most used commodities in the world, and it is one of the most extensively used fossil fuel [4] [46]. Therefore, developing a prediction tool to minimise the impact of this problem is paramount. Such a model needs to observe the possible factors that contribute to this volatility. Nelson, et al. in [13], hypothesises that the main factors affecting the market are oil demand and supply, populations, politics and economy. These main factors had been proven in [13] to be the key contributors to the volatility of the crude oil market based on a survey of geographically dispersed experts based on related economic variables.

This chapter will discuss the development of a hierarchical conceptual (HC) model as the basis platform for information exploration associated with the factors involved in crude oil price prediction, as well as the exploitation of these variables as possible inputs for a quantitative prediction model. Among the factors to be investigated in this model are demand, supply, politics, economy, production, consumption and population. Figure 3.1 exhibits the overall research framework with the parts to be discussed in this chapter highlighted. The development framework was inspired by the prediction development suggested by [7]. The novelty of this prediction models framework lies in the development of the HC model, rule-based expert model, linguistic prediction model and linguistic-quantitative prediction model, with linguistic information introduced into the models.

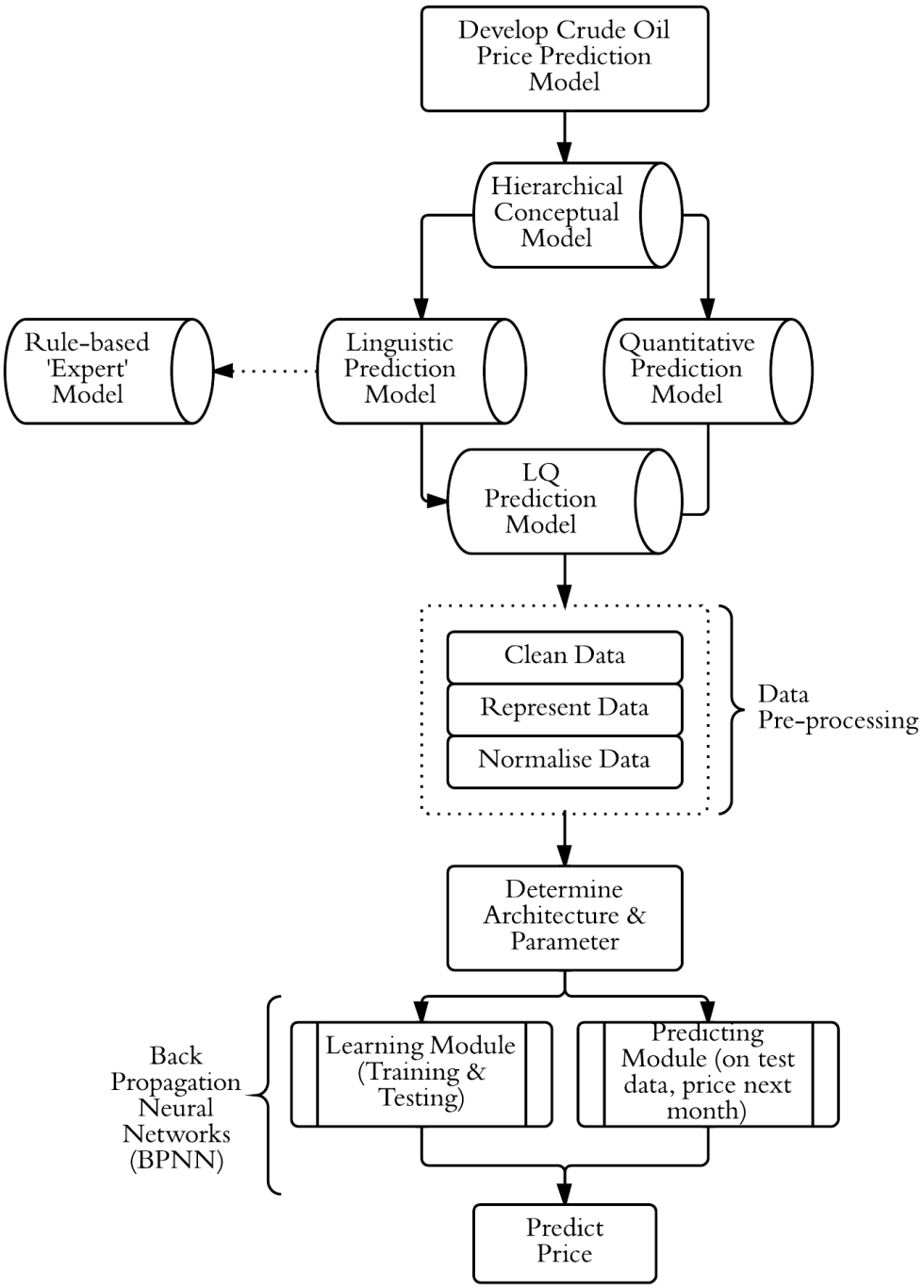


Figure 3.1 The Development Framework for Hierarchical Conceptual Model and Quantitative Prediction Model.

3.2 HIERARCHICAL CONCEPTUAL (HC) MODEL

To build a prediction model, factors contributing to the crude oil price volatilities need to be validated and verified to ensure the significance of these factors to be used as inputs in the prediction model. A hierarchical conceptual[87] (HC) model was introduced and developed in this research to fulfil this purpose to systematically identify and classify the possible related inputs. Online news is a great source of knowledge as it stores valuable information that can be potentially exploited as rules that affect the market. Through the HC model, a systematic approach based on online news mining was developed to retrieve the information relating to events containing the oil market impact factors via online news services. To comprehend the uncertainties involved in the market, factors that contribute to these uncertainties were first retrieved. The HC model, via its news mining approach, utilised such information to train a prediction network and store it into a database. The HC model was developed in this research through the processes of i) data classification, ii) information retrieval and iii) feature extraction as exhibited in Figure 3.2. These processes are explained in detail in the following subsections 3.2.1, 3.2.2, and 3.2.3.

[87] Hierarchical concept: a set of features arranged in a tree structure diagram, with each feature associated with instances of that feature.

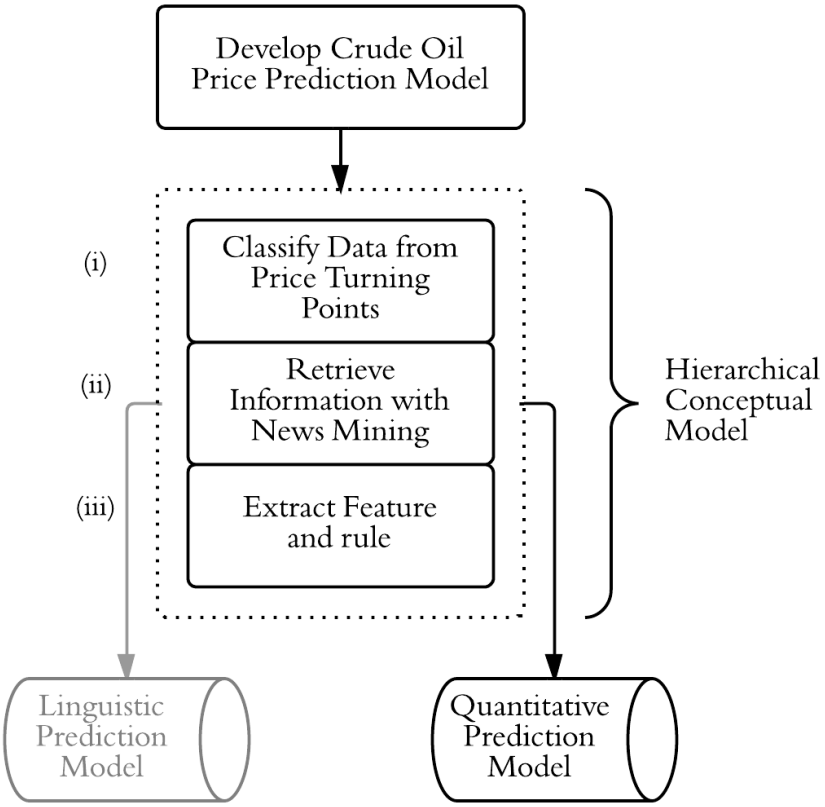


Figure 3.2 The Development Process of the Hierarchical Conceptual Model.

3.2.1 Data Classification Based on Price Turning Points

To develop a hierarchical conceptual (HC) model introduced in section 3.2, data were first prepared and segregated into different classes in order to retrieve relevant information from the online news service. The data classification process, which was conducted as a reference to mine the data from the online news articles, was prepared according to crude oil price turning points[89], referring to a particular period. Figure 3.3 shows an example of turning points occurring in the crude oil market.

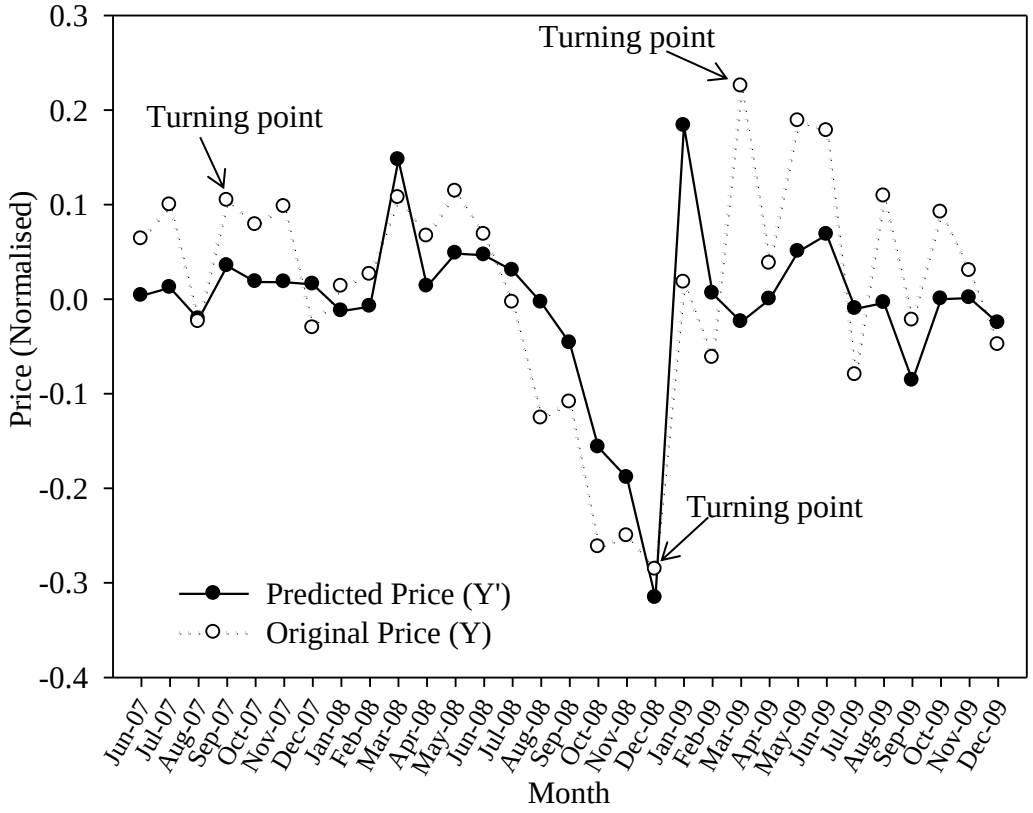


Figure 3.3 Monthly Crude Oil Price Turning Points for Jun 2007 to December 2009.

[89] Turning point: a point at which the crude oil price inverses direction. It tends to occur at a fixed rhythm time, such that every *X* duration there is a turning point for that price.

To retrieve information that corresponds to the volatility of the market, West Texas Intermediate (WTI) price data was used and classified into a) high oil price[91] and b) low oil price[92] classes. Each (a) and (b) classes were further classified into 3 categories: i) large impact; ii) medium impact; and iii) small impact, which classify the level of impact for events related to the crude oil price market. The classes for each (i) to (iii) were calculated based on equation (3.1) where, for each class, the price turning points were represented by its impact level as tabulated by Table 3.1.

$$T^{impact} = \left(\frac{x_2 - x_1}{x_1}\right) \text{ x } 100\%$$

where,

T^{impact} denotes the turning point impact;

x_2 denotes current price; and

x_1 denotes previous price. (3.1)

Table 3.1 Turning Points Impact Category Based on Data Classification Process for Hierarchical Conceptual (HC) Model.

CATEGORY	VARIANCE	
	HIGH OIL PRICE	LOW OIL PRICE
Large Impact	> 30%	> 20%
Medium Impact	20-30%	10-20%
Small Impact	< 20%	< 10%

In Table 3.1, the price turning points illustrated by Figure 3.3 were categorised as either a large, medium or small category based on monthly West Texas Intermediate (WTI) price data ranging from January 1984 to February 2009, with the calculations made via equation (3.1). For a high oil price case, a turning point carrying more than 30% variance is considered as having a large price impact, with 20-30% variance as having medium impact, and less than 20% as having low impact. Meanwhile, for the low oil price case, a turning point with more than 20%, 10-20% and less than 10% variance is considered as having large, medium and small impact of price,

[91] High oil price class: prices with upward direction.
[92] Low oil price class: prices with downward direction.

respectively. These impact classes were quantified into percentages based on equation (3.2) where the average value for each oil price cases was calculated.

$$A = \frac{1}{n}\sum(T^{impact}) \tag{3.2}$$

where,

A denotes the average value;

T^{impact} denotes the turning point impact calculated in equation (3.1);

and *n* as the total data.

The data classification process discussed in this subsection proposes a basis for further investigation into the events related to volatile prices. An event represented by the turning points will lead to a discovery of rules[95] to be implemented in the next models of this research in order to understand the behaviour of the crude oil market. The turning points categorised as having large and medium impact are considered significant as these indicate that significant events had occurred in the market. Hence, the key impact factors that contribute to the volatility of the market's price are contained in these turning points, and this is information to mine via the online news service. The large impact data were chosen as they contain sufficient information to discover the volatility activities in the market. News was mined via the online news service, with reference to the periods noted by the turning point prices in the large impact price category. Table 3.2 shows the example of data used for the classification of price impact. The shades of orange presented in the Table 3.2 denote the level of impact carried by each turning point, as calculated by equation (3.1). Further investigation into the behaviour of the crude oil price market continues in subsection 3.2.2 with the information retrieval process.

[95] Rule: a rule is defined as a set of explicit ascendances for an occurring event.

Table 3.2 Example on Classification of Price Impact in Excel
Based on Crude Oil Price Turning Points for High Oil Price and Low Oil Price Case

Date	Price x_1	Price x_2	%	High Price Turning Point			Date	Price x_1	Price x_2	%	Low Price Turning Point		
				Large	Medium	Small					Large	Medium	Small
Jan-85	25.64	28.81	0.1236				Mar-84	30.76	29.97	0.0257			
Jun-85	27.14	30.81	0.1352				Sep-84	29.31	25.64	0.1252			
Apr-86	12.85	15.44	0.2016				Apr-85	28.81	27.14	0.0580			
Jul-86	11.58	18.66	0.6114				Nov-85	30.81	12.85	0.5829			
Feb-87	17.73	21.36	0.2047				May-86	15.44	11.58	0.2500			
Mar-88	16.22	17.88	0.1023				Jan-87	18.66	17.73	0.0498			
Nov-88	13.98	17.98	0.2861				Jul-87	21.36	19.53	0.0857			
Feb-89	17.83	21.04	0.1800				Oct-87	19.85	16.22	0.1829			
Aug-89	18.52	19.82	0.0702				Apr-88	17.88	15.52	0.1320			
Nov-89	19.82	22.64	0.1423				Aug-88	15.52	13.98	0.0992			
Jul-90	18.64	35.92	0.9270				Apr-89	21.04	18.52	0.1198			
Jun-91	20.20	23.23	0.1500				Jan-90	22.64	18.64	0.1767			
Mar-92	18.92	22.38	0.1829				Oct-90	35.92	19.86	0.4471			
Mar-94	14.66	19.65	0.3404				Oct-91	23.23	18.92	0.1855			
Dec-94	17.16	18.55	0.0810				Jun-92	22.38	19.08	0.1475			
Mar-95	18.55	19.74	0.0642				Mar-93	20.35	17.87	0.1219			
Oct-95	17.44	19.04	0.0917				Oct-93	18.15	14.66	0.1923			
Feb-96	18.88	23.57	0.2484				Jul-94	19.65	17.16	0.1267			
Jun-96	20.45	24.90	0.2176				May-95	19.74	17.44	0.1165			
Nov-96	23.71	25.17	0.0616				Dec-95	19.04	18.88	0.0084			
Feb-99	12.01	17.89	0.4896				Apr-96	23.57	20.45	0.1324			
Jun-99	17.89	22.64	0.2655				Oct-96	24.90	23.71	0.0478			
Oct-99	22.64	29.89	0.3202				Jan-97	25.17	19.17	0.2384			
Apr-00	25.74	31.83	0.2366				Oct-97	21.26	15.44	0.2738			
Jul-00	29.77	34.40	0.1555				Apr-98	15.44	14.95	0.0317			
Feb-02	20.74	25.52	0.2305				Mar-00	29.89	25.74	0.1388			
Jun-02	25.52	26.27	0.0294				Jun-00	31.83	29.77	0.0647			
Nov-02	26.27	35.87	0.3654				Nov-00	34.40	28.46	0.1727			
May-03	28.14	31.59	0.1226				May-01	28.64	26.45	0.0765			
Sep-03	28.29	36.69	0.2969				Aug-01	27.47	20.74	0.2450			
Apr-04	36.69	40.28	0.0978				Feb-03	35.87	28.14	0.2155			
Jun-04	38.02	53.13	0.3974				Aug-03	31.59	28.29	0.1045			
Jan-05	46.84	54.31	0.1595				May-04	40.28	38.02	0.0561			
May-05	49.83	65.57	0.3159				Oct-04	53.13	46.84	0.1184			
Nov-05	58.30	61.63	0.0571				Mar-05	54.31	49.83	0.0825			
Feb-06	61.63	74.41	0.2074				Sep-05	65.57	58.30	0.1109			
Oct-06	58.88	62.03	0.0535				Jan-06	65.51	61.63	0.0592			
Jan-07	54.57	72.39	0.3266				Jul-06	74.41	58.88	0.2087			
Aug-07	72.39	94.62	0.3071				Dec-06	62.03	54.57	0.1203			
Jan-08	92.95	133.44	0.4356				Nov-07	94.62	92.95	0.0176			
Feb-09	39.16	59.16	0.5107				Jul-08	133.44	39.16	0.7065			
		average	0.2391						average	0.1575			
Large	Medium	Small					Large	Medium	Small				
> 30%	20-30%	< 20%					> 20%	10-20%	< 10%				

3.2.2 Information Retrieval

Data were utilised in the news mining process of this subsection, after the classification into categories of price impacts in subsection 3.2.1. The process in this subsection was to investigate the events involving the crude oil price market based on the large impact category. The impact factors contributing to the volatility of the crude oil market were also discovered by using the price turning points contained in the large impact category as a reference to mine the online news service. Another method used to retrieve information is text mining. Panas and Ninni [27] used text mining to retrieve information from the stock market and analyse its correlations. In this chapter, a news mining process via the Google News service was employed to retrieve such information. Google is well known as one of the biggest search engines in the world, and houses the Google News service. The Google News service's ability to extract and demonstrate the search results according to timelines is very helpful and was used in this research. More about Google News is discussed in section 4.2.1 of Chapter 4. Proper selections of keywords for news mining were chosen to return appropriate and relevant news. In order to download and mine data from the Google News service, the keywords 'crude oil: price' was used to retrieve information.

3.2.3 Feature and Rule Extraction

By utilising the extracted results from keyword searches from the news mining phase, features of the market were extracted and analysed to understand its behaviour. News is useful information to exploit as it offers knowledge that can be interpreted as patterns that can be then be amalgamated as rules. From the features extracted, rules (or patterns) were collected and stored in a database to be used in the quantitative prediction model discussed in section 3.3.

The features extracted from sentences in an article were based on the frequency of occurrences of a word i.e., a possible factor that gives impact to the crude oil prices (impact factor), amalgamated with a keyword: “crude oil price”; “crude oil”; or, “oil price” in a sentence of an article; hence extracted features = sentence (impact factor + keyword). For example, an article with this sentence: “Crude oil trades near USD70 a barrel on demand concern in China” indicates “Crude oil” as the keyword found, with “demand” as the impact factor. An impact factor found in a sentence is selected as a feature based on the frequency with which it occurs in the online news article. The impact factors used and discussed in [13] were also used as a reference to extract these features. Discussions on crude oil price behaviour in an article, especially those with drastic after effects, are often associated with the factors that caused it, in a cause-effect relationship. From this feature-extraction process, 1,139 monthly news articles were mined from the large impact category and were later amalgamated into the rule extraction phase. The manual rule extracting process was used to discover the factors that contributed to the volatility of the market. The impact factors discovered from this rule extraction are presented in Figure 3.4.

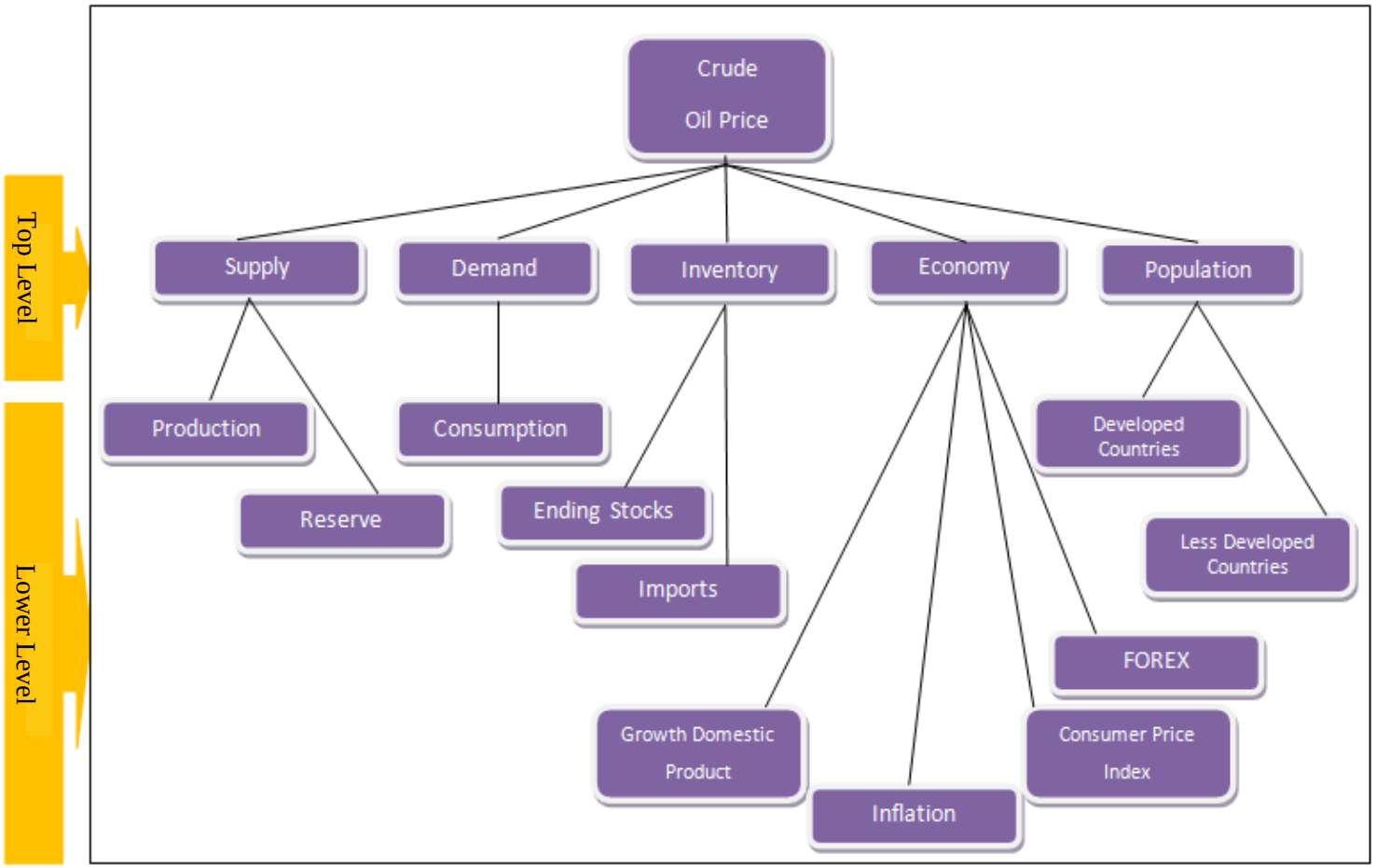


Figure 3.4 The Key Impact Factors Contributing to Crude Oil Price Volatility Based on Hierarchical Conceptual (HC) Model.

Based on the frequency of occurrences, the features extracted from the articles were classified into hierarchical conceptual information with i) the top level representing the core contributor and ii) lower level as the subordinate contributor. This hierarchy information will provide a better understanding of market behaviour and its contribution to the volatilities. The factors discovered from the extracted features were tested in a quantitative prediction model as discussed in subsections 3.3.4.1 and 3.3.4.2 to systematically examine and the evidence of its correlations with the crude oil price market. It was hypothesised that fairly good prediction outcomes are anticipated, according to the convergence produced from the quantitative prediction model.

Based on the three processes mentioned in subsection 3.2.1, 3.2.2 and 3.2.3, 22 quantitative data variables denoted as the main contributors to the crude oil price market were extracted and are presented as in Table 3.3. These quantitative features were used as the input in the quantitative model discussed in the next section 3.3, wherein its credibility to be exploited as an input was tested.

Table 3.3 The Key Impact Factors for Crude Oil Price Discovered from HC Model and used in the Quantitative Prediction Model.

NO.	VARIABLES	FACTORS
	$\boldsymbol{S^{T}}$[97]	**SUPPLY**
1	S_{a1}	Productions of OPEC countries
2	S_{a2}	Productions of Non-OPEC countries
3	S_{b1}	Proved reserves of OPEC countries
4	S_{b2}	Proved reserves of OECD countries
5	S_{b3}	Number of well drilled
	$\boldsymbol{D^{T}}$	**DEMAND**
6	D_{a1}	Consumption of OECD countries
7	D_{a2}	Consumption of China
8	D_{a3}	Consumption of India
	$\boldsymbol{I^{T}}$	**INVENTORY**
9	I_{a1}	Ending stocks of OECD countries
10	I_{a2}	Ending stocks of US
11	I_{b1}	US petroleum imports from OPEC countries
12	I_{b2}	US petroleum imports from Non-OPEC countries
13	I_{c1}	US crude oil imports from OPEC countries
14	I_{c2}	US crude oil imports from Non-OPEC countries
	$\boldsymbol{E^{T}}$	**ECONOMY**
15	E_{a1}	Foreign Exchange of GBP/USD
16	E_{a2}	Foreign Exchange of Yen/USD
17	E_{a3}	Foreign Exchange of Euro/USD
18	E_{b1}	US Growth Domestic Products (GDP)
19	E_{c1}	US Inflation rate
20	E_{d1}	US Consumer Price Index (CPI)
	$\boldsymbol{P^{T}}$	**POPULATIONS**
21	P_{a1}	Population of developed countries
22	P_{a2}	Population of less developed countries
23	***WTI***	**WEST TEXAS INTERMEDIATE PRICE**

[97] T is the accumulated value or the core contributor of each feature.

3.3 ARTIFICIAL NEURAL NETWORKS (ANN) FOR THE QUANTITATIVE (ANN-Q) PREDICTION MODEL

ANN has gained much attention for its computational intelligence approach and its capability to make predictions. Its capability of modelling nonlinearity results in a class of general function approximations [47]. The development of this quantitative prediction model was adapted from a process development suggested by Rao and Rao [48] and is presented in Figure 3.5 of this section. In [48], the authors suggest a systematic development approach to forecasting the stock market using the back propagation neural network (BPNN). The authors emphasise the importance of choosing the right inputs for network training; if a relationship between the data is weak, the network will learn to ignore it automatically. The prediction development in this chapter consists of three phases: i) objective determination, ii) data pre-processing and (iii) ANN modelling of each is discussed in subsections 3.3.1, 3.3.2 and 3.3.3, respectively and presented by Figure 3.5.

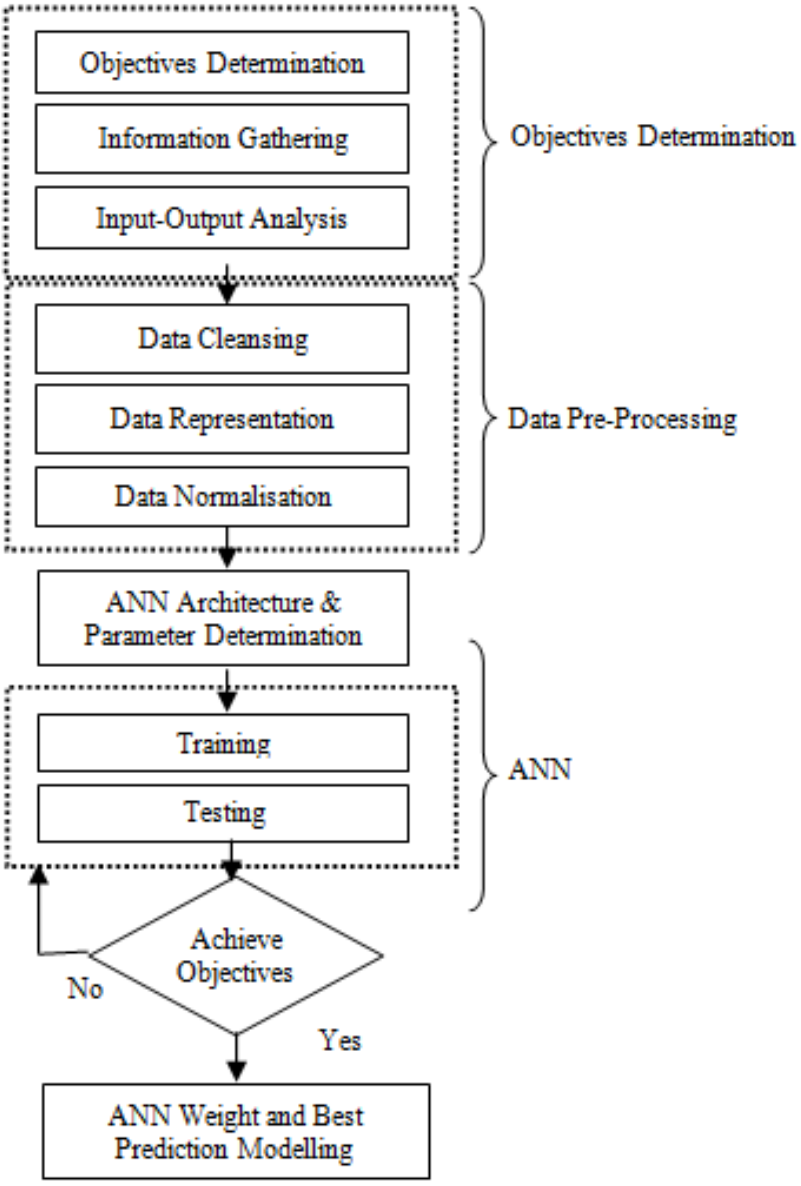


Figure 3.5 Artificial Neural Networks (ANN) Development Framework [48].

3.3.1 Objective Determination

The framework development of this quantitative prediction model starts with determining the model's objectives. The development of the quantitative prediction model focuses on building a robust and accurate prediction tool for the crude oil market based on monthly quantitative data, and predicts its price for every barrel in US Dollars (USD). Based on the objective, vectors of inputs and output were carefully chosen to fit into the network. Appropriate selection of inputs for learning is essential for the network to propagate connections effectively. Failing in the selection of the appropriate feature set will lead to a reduction of the prediction performance which demonstrates the instability of the network.

Based on information retrieved from the HC model, a period of 25 years was chosen as a basis to collect the quantitative data for the quantitative prediction model. The collected data range from January 1984 to February 2009 were sourced from Economagic [8], Energy Information Administration (EIA) [10], World Bank [14], International Energy Agency (IEA) [49], World Energy [50], Population Reference Buraeu [51], , and Source OECD [52]. This data were then processed and normalised in order to removc noise and missing values.

3.3.2 Data Pre-processing

In this data pre-processing phase, the data collected from sources [8] to [14] mentioned in subsection 3.3.1 were given an extensive consideration to ensure that they were normalised and noise reduction was performed. This data pre-processing process is also established to ensure that only data from reliable and premiere sources were used for the prediction model, as suggested by Nasrudin [53]. Collecting data from related official agencies and presenting them as a time-series will avoid missing or inappropriate values of data. The data pre-processing process was aimed to achieve fitter data sets to be induced in the network for the quantitative prediction model. In the quantitative prediction model, the pre-processed data were then represented and simulated based on two different types of data: time-series and normalised. To gain normalised data, the collected time-series data were normalised using the One-Step Returns function[98] as presented as in equation (3.3) where *R*

[98] Suggested by [20].

denotes the current month returns value for input number *n* with X_n and X_{n-1}denote as the current and previous input respectively.

$$R_n = \frac{X_n - X_{n-1}}{X_{n-1}} \tag{3.3}$$

By using equation (3.3), the data for the quantitative prediction model were normalised to produce values between -1 and 1. Table 3.4 shows an example of a database used for the quantitative prediction model which is presented in normalised form. These two different types of data representation (time-series and normalised) were used for simulating the validation of the compatibility and the suitability of both types in representing the data for prediction.

Table 3.4 Example of Data Represented in Normalised Form for the Quantitative Prediction Model.

Year	Month	Supply					Demand		
		Production		Reserves			Consumption		
		OPEC	Non-OPEC	Well Drilled	Proved Reserves OPEC	Proved Reserves OECD	OECD	China	India
1984	January	-0.040	0.022	0.507	0.009	-0.005	0.047	0.228	-0.206
	February	0.002	0.008	-0.089	0.000	0.000	0.009	0.000	0.000
	March	0.016	-0.015	0.103	0.000	0.000	-0.007	0.000	0.000
	April	-0.007	0.009	-0.302	0.000	0.000	-0.074	0.000	0.000
	May	-0.036	0.006	0.313	0.000	0.000	0.003	0.000	0.000
	June	0.078	0.000	-0.046	0.000	0.000	-0.006	0.000	0.000
	July	-0.050	0.004	-0.021	0.000	0.000	-0.007	0.000	0.000
	August	-0.071	-0.013	0.000	0.000	0.000	0.029	0.000	0.000
	September	0.010	0.008	0.141	0.000	0.000	-0.024	0.000	0.000
	October	0.002	0.010	-0.286	0.000	0.000	0.033	0.000	0.000
	November	-0.003	0.002	0.200	0.000	0.000	0.028	0.000	0.000
	December	-0.013	0.004	-0.211	0.000	0.000	-0.005	0.000	0.000
1985	January	-0.052	-0.011	0.169	0.065	0.023	0.053	0.083	-0.039
	February	0.078	0.008	-0.012	0.000	0.000	0.000	0.000	0.000
	March	-0.002	0.006	0.067	0.000	0.000	-0.080	0.000	0.000
	April	-0.026	0.000	-0.286	0.000	0.000	-0.028	0.000	0.000
	May	-0.087	-0.002	0.192	0.000	0.000	-0.018	0.000	0.000
	June	-0.042	-0.022	-0.201	0.000	0.000	-0.006	0.000	0.000
	July	0.027	0.017	0.151	0.000	0.000	0.031	0.000	0.000
	August	0.002	-0.004	0.212	0.000	0.000	0.017	0.000	0.000
	September	0.091	0.016	-0.301	0.000	0.000	-0.021	0.000	0.000
	October	0.100	0.004	0.138	0.000	0.000	0.050	0.000	0.000
	November	0.020	0.003	-0.295	0.000	0.000	-0.004	0.000	0.000
	December	0.032	-0.006	0.484	0.000	0.000	0.051	0.000	0.000

3.3.3 Back Propagation (BP) Trained Neural Network (BPNN) Modelling

Artificial Neural Networks (ANN) using the Back Propagation technique was employed in this model as the learning algorithm to learn from the input data. Further discussion on this algorithm is thoroughly documented in Chapter 5 of this thesis. BPNN is known to have the ability to map complex non-linear functions, as it contains multiple layers of function-mapping neurons [41]. The networks were trained by evolving the weigh values using temporal input data to adjust and later compare the weighted values with a threshold value. The network is activated to obtain satisfactory error criteria from the network, if the net input is greater than or equal to the threshold value. The sample architecture of this network is described in Table 3.5 of subsection 3.3.4.

3.3.4 Network Architecture and Parameters

To activate the model, a set of parameters was chosen to optimise the process of training the artificial neural networks (ANN) and to minimise the errors for accurate prediction outcomes. The parameters utilised for this network model are shown in Table 3.5 where these parameters were utilised by the learning algorithm discussed in subsections 3.3.4.2 and 3.3.4.3. In most instances, the literature suggests the use of a trial-and-error approach in determining network parameters, where the performance goal is set by the user. Amongst the suggestions are from [39] [54]. Wang, et al. [39] and Rene, et al. [54] used the trial-and-error approach to configure the network parameters for their simulations. The parameter values for initial weight, learning rate, learning cycle and momentum of the network discussed in this thesis were also chosen and set by trial and error, based on the findings made and discussed by [55] [56], where their parameter values were used as reference and trial. By keeping some training parameters constant and slowly moving the other parameters over an adequate range of value as suggested by [56] [57] [58], the final parameters induced in the quantitative prediction model are tabulated in Table 3.5. Zero-based log sigmoid, log sigmoid and hyperbolic tangent were used as the activation function for the computing element in the network. A sensitivity analysis was taken in section 3.4 to evaluate the performance of these parameters and to ensure the parameters chosen are optimal.

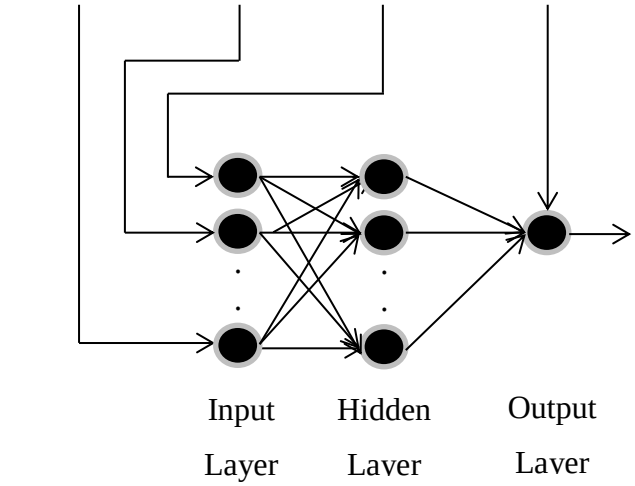
Table 3.5 Back Propagation Neural Network (BPNN) Parameters for Quantitative Prediction Model.

PARAMETER	**VALUE**
Initial Weight	0.3
Learning Rate	0.3
Momentum	0.6
Learning Cycle (epochs)	5000
Network Layer	3
Neurons:	
Input Layer	22
Hidden Layer	3
	4
	5
Output Layer	1
Activation Function	Zero-based Log Sigmoid
	Log Sigmoid
	Hyperbolic Tangent

Table 3.6 presents an example of a network feed forward phase for a three-layer BPNN where the input and output data selected are a vector induced to the input nodes and the target is used to train the output layer. In BPNN, inputs are propagated throughout the network in a sequence through the input-hidden-output layers. Here, the hidden layer translates the input activation through sigmoidal neurons to the output layer and, later in the training phase, back-propagates the σ (error) signals from the output back to the input layer.

Table 3.6 Example of Artificial Neural Networks (ANN) Architecture For Crude Oil Price Prediction

INPUT x_1	**INPUT** $x_{..}$	**INPUT** x_n	**OUTPUT (TARGET PRICE)** y_{t1}
x_{11}	$x_{..1}$	x_{n1}	y_{t11}
$x_{1..}$	$x_{....}$	$x_{n..}$	$y_{t1..}$
x_{1n}	$x_{..n}$	x_{mn}	y_{t1n}
.	.	.	.



Single-digit hidden neurons such as 3, 4 and 5 were depicted in the network to avoid computational burden and to maintain network propagating stability. A sensitivity analysis via the trial-and-error method was employed in order to analyse the generalisation accuracy. Nevertheless, different methods can be used for the analysis as suggested by [26]. The learning module and the predicting module are introduced in the subsections 3.3.4.1 and 3.3.4.2, respectively. These modules governed the core process development of the quantitative prediction model.

3.3.4.1 The Learning Module

The monthly time-series quantitative data used in this chapter were derived from features discovered in the hierarchical conceptual (HC) model in section 3.2. The quantitative data employed in the network represent the key input factors that contribute to the volatility of the crude oil price market, as tabulated in Table 3.7.

Table 3.7 Sensitivity Analysis for the Quantitative Data Employed in Back Propagation Neural Networks (BPNN)

TRAINING NO.	DATA SELECTION	EXCEPTIONAL DATA	ATTRIBUTE NO.
1	All data	N/A	22
2	All except I_T[99]	I_{a1},I_{a2}, I_{b1}, I_{b2}, I_{c1}, I_{c2}	16
3	All except E_T	E_{a1}, E_{a2}, E_{a3}, E_{b1}, E_{c1}, E_{d1}	16
4	All except P_T	P_{a1}, P_{a2}	20
5	All except (I_T + E_T)	(I_{a1},I_{a2}, I_{b1}, I_{b2}, I_{c1}, I_{c2}) + (E_{a1}, E_{a2}, E_{a3}, E_{b1}, E_{c1}, E_{d1})	10
6	All except (I_T + P_T)	(I_{a1},I_{a2}, I_{b1}, I_{b2}, I_{c1}, I_{c2}) + (P_{a1}, P_{a2})	14
7	All except (E_T + P_T)	(E_{a1}, E_{a2}, E_{a3}, E_{b1}, E_{c1}, E_{d1}) + (P_{a1}, P_{a2})	14
8	All except (I_T + E_T + P_T)	(I_{a1},I_{a2}, I_{b1}, I_{b2}, I_{c1}, I_{c2}) + (E_{a1}, E_{a2}, E_{a3}, E_{b1}, E_{c1}, E_{d1}) + (P_{a1}, P_{a2})	8

[99] I_T, E_T and P_T refer to Table 3.3.

The data used in the quantitative prediction model consisted of 302 vectors, each composed of 23 scalars (22 inputs, 1 output) and the 6,946 temporal vectors[100] ranging from January 1984 to February 2009. These observations were derived from 1,139 news articles mined initially and derived from the HC model. For learning, a sensitivity analysis was made to evaluate the total prepared data. Total data were divided into a training: testing percentage ratio of 90:10 percent, 80:20 percent, and 70:30 percent[101] and trained individually with 3, 4 and 5 hidden neurons.

The optimal training set was determined based on the minimum value of absolute error[102] as produced by the network. In the learning module, the data were forward-propagated through a sequence of input-hidden-output layer and was prepared and represented as i) one-step returns function and ii) time-series. These sets of data were also trained with eight different vectors of input variables as presented in Table 3.7. The low-error value produced indicates the fully connected[103] layers in the network. This process validates the key factors discovered in the HC model with the hypothesis that the dataset with the lowest errors represents the data optimally interconnected in the network.

[100] Temporal vectors: sequence of training inputs.
[101] 90: 10, 80:20 and 70: 30 ratios: a training set with 90 percent of total data used for training and the remaining 10 percent for testing, or a training set with 80 percent of total data used for training and the remaining 20 percent for testing, or a training set with 70 percent of total data used for training and the remaining 30 percent for testing.
[102] Absolute error: the magnitude difference of between the exact value and the approximation.
[103] Fully connected: every neuron in each layer is connected to every other neuron in the adjacent forward layer.

3.3.4.2 The Predicting Module

For the network discussed in subsection 3.3.4.1, learning from the training set of input pattern provided by the quantitative information obtained from the HC model was depicted in the network. The network computes its output pattern by adjusting the weights to achieve a desired output pattern by iterating to minimise the error. In this module, independent sets of data were injected into the network to test its predicting capability. This is formulated by equation (3.4) and equation (3.5) with y_{n+1} denoting the future outcome with:

$$y_{n+1} = f\left(\sum_{i=1}^{n} x_1 w_{11} + x_2 w_{21} + x..w_{..1} + x_n w_{m1}\right) - \theta_1 \tag{3.4}$$

where, f (.) represents the sigmoid function for the first month period of prediction with x, w and $\theta_1$denoting the input, weight and error (threshold) to the neuron respectively. The general equations for the future price were represented by equation (3.5) and (3.6). The variables were defined in a nomenclature[108] and presented in Table 3.7.

$$WTI_{n+1} = f(S^T + D^T + I^T + E^T + P^T) - \theta_n \tag{3.5}$$

$$WTI_{n+1} = f\left[\sum_{n=1} (S_{a1+a2+b1+b2+c1}) + (D_{a1+a2}) + (I_{a1+a2+b1+b2+c1+c2}) + (E_{a1+a2+a3+b1+c1+d1}) + (P_{a1+a2})\right] - \theta_n \tag{3.6}$$

The outcomes from this prediction are evaluated by three evaluation metrics to validate the network credibility of the quantitative prediction model to predict the price of the crude oil market.

[108] Nomenclature: a system of names or terms, or the rules for forming these terms in a particular field.

3.3.4.3 Performance Evaluation Metrics

The sum of squared errors is a useful indicator to evaluate a network performance. A network is considered converged when the value of the sum squared errors is sufficiently small. The root mean squared error (RMSE), normalised mean squared error (NMSE) and directional change statistic (D_{stat}) were employed to measure and validate the performance of this quantitative prediction model. The values of RMSE and NMSE through MSE were formulated based on equations (3.7), (3.8) and (3.9). While the sum of squared errors as the network performance indicator is useful, a directional performance indicator is an expedient measurement to determine the correctness of a directional prediction of an output pattern generated by the network. Hence, an additional performance metric is introduced and computed by equation (3.10).

$$MSE = \frac{1}{N}\sum_{t=1}^{N}(y_t - \hat{y}_t)^2 \tag{3.7}$$

$$NMSE = \frac{MSE}{\frac{1}{N}\sum(\hat{y}_t - A)^2} \tag{3.8}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(y_t - \hat{y}_t)^2} \tag{3.9}$$

$$D_{stat} = \frac{1}{N}\sum_{t=1}^{N} a_t \tag{3.10}$$

By referring to equations (3.7) to (3.10), *N* denotes the numbers of evaluation inputs, where y_t and $\hat{y}_t$ denote the actual and target predicted output respectively for time *t* with *A* as the average value for the absolute error. Nevertheless, *a* and *t* in equation (3.10) is expressed by, $a = 1$, if $(y_{t+1} - y_t)(\hat{y}_{t+1} - y_t) \geq 0$ and a $= 0$ otherwise.

The sensitivity of the network architecture chosen is analysed in section 3.4. The sensitivity analysis process, which is also known as a what-if analysis, is a simulation analysis in which the key quantitative, assumptions and computations (underlying a decision, estimate, or project) are changed systematically to assess their effect on the final outcome. The results of the sensitivity analysis were evaluated by equations (3.7) to (3.10) in the next section.

3.4 SIMULATION TESTS AND RESULTS

A series of datasets employed for training were applied to the network to analyse the sensitivity of the network in learning from different sets of data. The data represented in time series and normalised data were both trained and compared. The minimum value of the absolute error generated from these trainings was selected as the input data.

3.4.1 Sensitivity Analysis: Network Architecture and Parameters

The data were divided into eight sets of data training with different input vectors and different number of neurons. These training datasets were conducted in order to choose the best trained dataset in the learning module, as presented by the minimum absolute error value, to be employed in the propagation network. The training datasets were divided into three different sets of "training: testing" data ratios: i) 90:10 percent data ratio; ii) 80:20 percent data ratio; and iii) 70:30 percent data ratio. Each of these datasets was trained with three different sets of hidden neurons: i) 3 hidden neurons; ii) 4 hidden neurons; and iii) 5 hidden neurons. Table 3.8 shows the example of data represented in time-series with recorded actual vs. target predicted output results.

Table 3.8 Example of Actual Data vs Target Predicted Data Represented in Time Series with Hidden Neurons = 4

Hidden Nodes 4								
70:30			80:20			90:10		
Absolute Error	Original Data	Output Result	Absolute Error	Original Data	Output Result	Absolute Error	Original Data	Output Result
9.27704	26.270	16.47855	8.33156	34.740	33.78254	0.08432	73.050	69.751
	29.420	18.04496		36.760	33.04531		63.870	67.73108
	32.940	16.71026		36.690	33.12541		58.880	57.82326
	35.870	17.71045		40.280	33.59432		59.370	68.01387
	33.550	18.38483		38.020	33.53725		62.030	65.36423
	28.250	16.52721		40.690	31.21363		54.570	64.7493
	28.140	17.34007		44.940	32.00045		59.260	63.23969
	30.720	16.27243		45.950	31.78485		60.560	60.2579
	30.760	16.78655		53.130	33.0023		63.970	67.39491
	31.590	17.00693		48.460	34.04135		63.460	67.25271
	28.290	17.17466		43.330	33.66928		67.480	65.4281
	30.330	15.34588		46.840	31.51626		74.180	68.8056

As the performance of a network depends on the input that it received, an appropriate and good training data represented by the minimum absolute error is hypothesised to contain appropriate interconnections among the neurons. Good training data presented to the network will result in a more accurate targeted outcome. To prove the hypothesis, the data were divided into sets (i) to (iii). Table 3.9 shows the results from the sensitivity analysis conducted with training sets (i) to (iii).

Table 3.9 The Absolute Error Results for Training Sets with Hidden Neurons = 3, 4 and 5, and Trained with Normalised 90:10 percent, 80:20 percent and 70:30 percent of Training: testing Data Ratio for the Quantitative Prediction Model.

TRAINING SET NO.	**HIDDEN NEURON**	**TRAINING: TESTING DATA RATIO**		
		70:30	**80:20**	**90:10**
(i)	3	0.21691	4.23868	4.82512
(ii)	4	0.18955	2.11573	5.77583
(iii)	5	0.19488	0.03586	7.51819

Table 3.9 shows the best minimum absolute values produced from each training sets (i) to (iii). From the training, the minimum absolute value was converged from the training set (iii), with 0.03586 of absolute value. The network in training set (iii) was trained with 80:20 percent ratio of total data and 5 hidden neurons.

Table 3.10 presents the analysis made to measure the sensitivity of input data chosen and prepared for the network. The sensitivity analysis of this input data was made to validate the features extracted from HC model and also to determine the best input data to be applied to the network. The analysis started by preparing the network to train eight different training sets:

1) the training set with the application of all extracted features derived from HC model;
2) the training set with the application of all features except the inventory, *I* factors;
3) the training sets with the application of all features except the economy, *E* factors;
4) the training sets with the application of all features except the population, *P* factors;
5) the training sets with the application of all features except the inventory and economy factors;
6) the training sets with the application of all features except the inventory and population factors;
7) the training sets with the application of all features except the economy and population factors; and finally,
8) the training sets with the application of all features except the inventory, economy and population factors.

Table 3.10 The Results from Sensitivity Analysis for the Normalised Quantitative Data Employed in Back Propagation Neural Networks (BPNN) with 80:20 Percent Ratio of Total Data and Hidden Neuron = 5.

TRAINING SET	INPUT	EXCEPTIONAL INPUT	TOTAL NUMBER OF INPUT SCALAR	ABSOLUTE ERROR
1	All data	N/A	22	0.03586
2	All except I_T[110]	I_{a1},I_{a2}, I_{b1}, I_{b2}, I_{c1}, I_{c2}	16	2.73895
3	All except E_T	E_{a1}, E_{a2}, E_{a3}, E_{b1}, E_{c1}, E_{d1}	16	2.98231
4	All except P_T	P_{a1}, P_{a2}	20	5.20779
5	All except (I_T + E_T)	(I_{a1},I_{a2}, I_{b1}, I_{b2}, I_{c1}, I_{c2}) + (E_{a1}, E_{a2}, E_{a3}, E_{b1}, E_{c1}, E_{d1})	10	2.85359
6	All except (I_T + P_T)	(I_{a1},I_{a2}, I_{b1}, I_{b2}, I_{c1}, I_{c2}) + (P_{a1}, P_{a2})	14	1.18130
7	All except (E_T + P_T)	(E_{a1}, E_{a2}, E_{a3}, E_{b1}, E_{c1}, E_{d1}) + (P_{a1}, P_{a2})	14	2.25965
8	All except (I_T + E_T + P_T)	(I_{a1},I_{a2}, I_{b1}, I_{b2}, I_{c1}, I_{c2}) + (E_{a1}, E_{a2}, E_{a3}, E_{b1}, E_{c1}, E_{d1}) + (P_{a1}, P_{a2})	8	5.57810

From Table 3.10, the minimum absolute error value of 0.03586 was derived from training set 1 via 5,000 epochs of training. Training set 1, the largest and the total input used in the quantitative prediction model, consisted of 22 input data. This optimal result was followed by training sets 6 and 7 which exclude inputs *I* and *E* together with *P* from the training. Through the observation made on the training sets 1 to 8, inputs *I, E* and *P*, denoting the inventory, economy and population factors respectively, had shown significant interconnections and are important input neurons to the network. Training set 8 had produced the maximum value of 5.5781 absolute error, with the absence of these three input factors, which shows that, for the network to converge with minimum value of absolute error, an interconnection with either one or all of these *I, E* and *P* input is needed.

[110] I_T, E_T and P_T refer to Table 3.3.

Based on the results obtained from this simulation, the best input data to be used in the network are discovered. The promising 0.03586 absolute error value converged by training set 1 is validated as a significant preliminary evidence to the successful features discovered in the HC model. Training set 1 was utilised in the network with its sensitivity to the parameters depicted in subsection 3.3.4 and was analysed and tabulated in Table 3.11.

Table 3.11 The Performance Results based on the Sensitivity Analysis Made with the Parameters Depicted in Subsection 3.3.4, Trained with Normalised Data and Hidden Neurons = 5.

PERFORMANCE INDICATOR	**TRAINING: TESTING DATA RATIO**		
	70:30	**80:20**	**90:10**
NMSE	1.10700	0.00896	7.57680
Dstat (%)	77.66	93.33	70.00
RMSE	0.10100	2.26900	0.13590

The sensitivity analysis of this quantitative prediction model is indicated by normalised mean squared error (NMSE) as the primary performance indicator. The directional statistics (D_{stat}) is referred to as the secondary, followed by root mean squared value (RMSE).

Based on Table 3.11, the best prediction performance was derived from the training set that was trained with the allocation of 80 percent data used for training and the remaining 20 percent of total data used for testing. The training set shares promising values of 0.00896 NMSE, 93.33% for D_{stat}, and, finally, 2.2690 for RMSE. The prediction was made based on a testing data ranging from March 2004 to February 2009 and visualised in Figure 3.6.

Figure 3.6 and Figure 3.7 show the error when comparing the actual and predicted price, which visualises the accuracy of the quantitative prediction model. The accurateness shows not just the temporal local volatility but also the accuracy of its

discrete value. Therefore, the result from the quantitative prediction model proves and validates the variables chosen from the features extracted by hierarchical conceptual (HC) model. In addition, the directional gradient showed by the predicted price in Figure 3.6 was constantly positive and parallel[111] against the actual price. Furthermore, the Figure also shows strong interconnections held by the factors in the network recorded in Table 3.10 for training set 1.

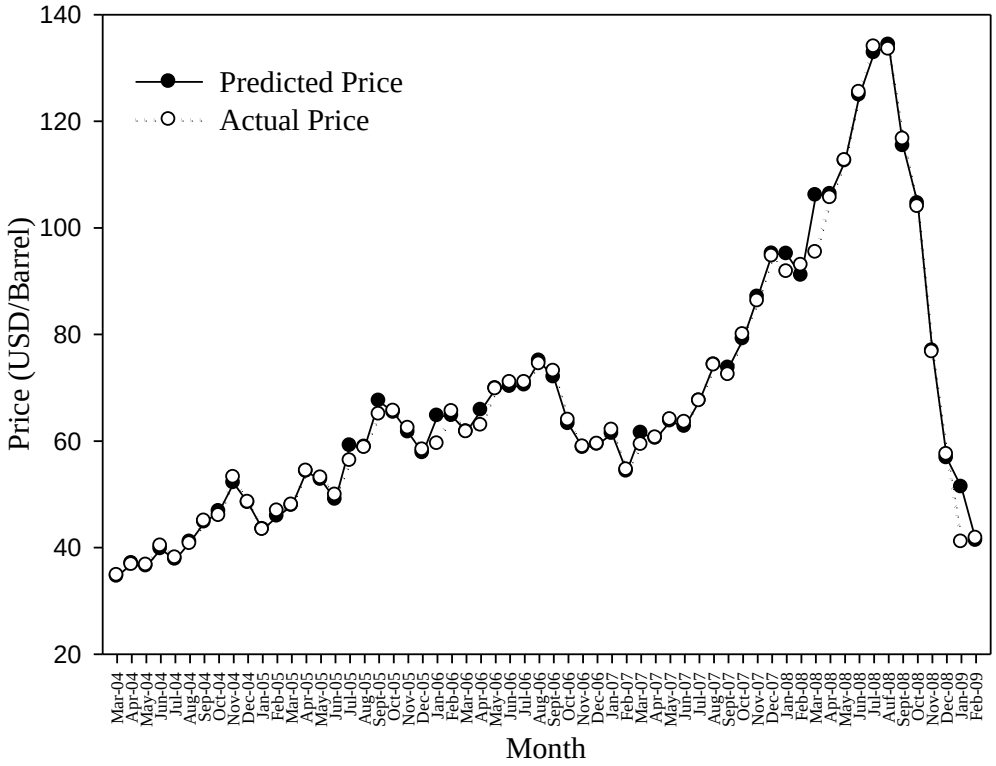


Figure 3.6 The Quantitative Prediction Model Optimal Performance Based on 80:20 Percent Training: testing Data Ratio with Hidden Neurons = 5.

[111] Positive and parallel: the gradients of two different variables are considered to be positive and parallel when moving in the same direction.

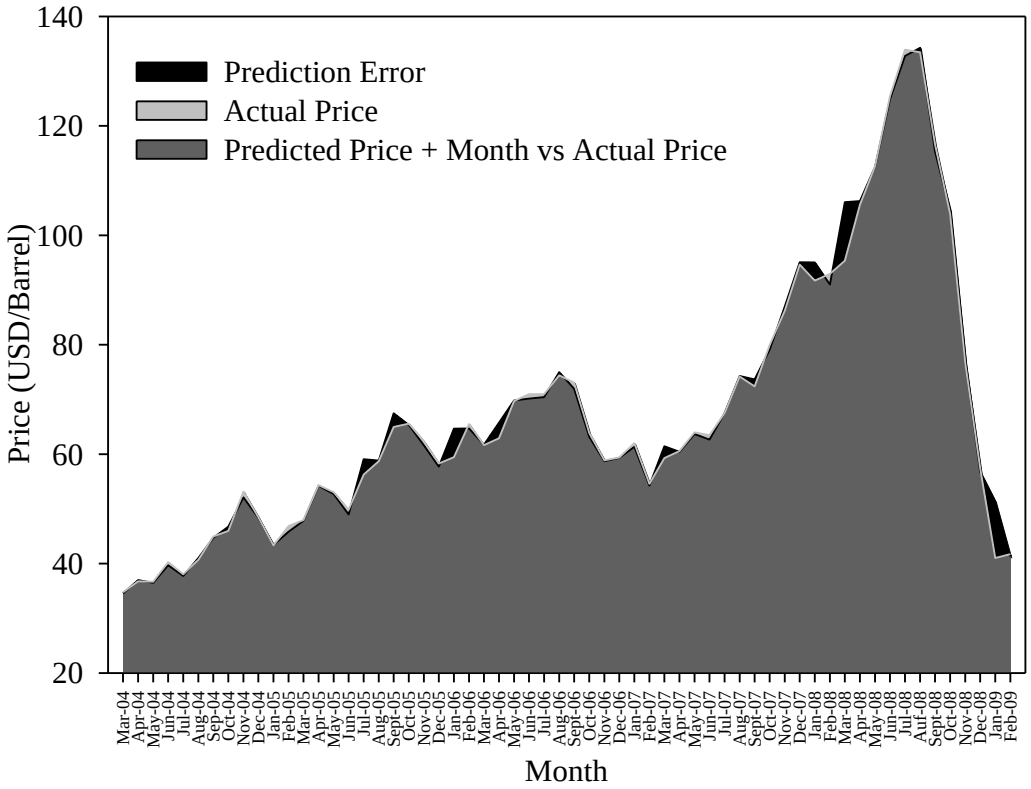


Figure 3.7 The Error Mapping of the Quantitative Prediction Model Based on 80:20 Percent Training: testing Data Ratio with Hidden Neurons = 5.

Section 3.4.2 will discuss the comparison of this quantitative prediction model with other prediction models used as the benchmark for the quantitative prediction model evaluation. The benchmark models were selected based on the application of back propagation neural networks (BPNN) as the learning algorithm, and the implementation of different techniques (hybrid) to derive a better crude oil price prediction model. Good benchmarks for evaluating the performance of the hybrid model are discussed later in Chapter 5.

3.4.2 A Comparison Analysis of Artificial Neural Network-Quantitative (ANN-Q) Prediction Model with Other Crude Oil Price Prediction Models

There is only very limited research on oil price forecasting, including quantitative and qualitative methods. As a single prediction model, an artificial neural network-quantitative (ANN-Q) model with the prospect of developing a hybrid linguistic-quantitative prediction model was compared and analysed against two hybrid models as performance benchmarks. The two models compared were the TEI@I Nonlinear Integration Forecasting model [39] and the EMD-FNN-ALNN model [43]. Both models used back propagation neural networks (BPNN) as their function mapping tool. The evaluation of the models with ANN-Q model was conducted based on the minimum value of root mean squared error (RMSE) as the primary error metric benchmarks, and the maximum value of directional statistics (D_{stat}) as the secondary. The analysis made was based on the performance of the three models. A single and two hybrid prediction models show that the utilisation of data chosen in ANN-Q model derived from the extracted features discussed in HC model demonstrates competitive prediction results with 93.33% of directional accuracy presented by the $D_{stat.}$; the second-highest accuracy after model [39] with 95.83% and model [43] with 86.99%. Nonetheless, ANN-Q performed last for its 2.2690 RMSE value, compared to model [39] and [43] with 1.0549 and 0.2730 of RMSE value, respectively. Even with a maximum value of error, ANN-Q is a promising tool for prediction given its directional accuracy. The accurate directional performance fairly interprets the strong interconnections of input data held by the network for use in the prediction, even as a single prediction model. ANN-Q would have performed better than the other two models if compared to its normalised mean squared error (NMSE) result.

The evaluation analysis also validates the importance of data pre-processing in preparing the data before learning, which was absent in the other compared models. According to the RMSE value owned by EMD-FNN-ALNN, the decomposition[113] and ensemble (or 'divide' and 'conquer') techniques do support its claim as an alternative prediction tool for the crude oil market. Nonetheless, in terms of D_{stat} comparison, *TEI@*I Nonlinear Integration has outperformed the other two models with 95.83% directional accuracy, followed by ANN-Q with 93.33% and, lastly,

[113] Decomposition: the process of breaking down substances into simpler form.

EMD-FNN-ALNN with 86.99%. Even though ANN-Q performed last for its RMSE value, the maximum RMSE does not necessarily mean that there is a high hit rate in predicting the crude oil price movement [43]. Hence, D_{stat} is more practical in reflecting the fluctuating trend of the WTI price. However, ANN-Q is still a promising prediction tool for crude oil price with only 2.5% less than the D_{stat} winner, the TEI@I Nonlinear Integration model. Albeit being a single model prediction, ANN-Q proved to be competitive and comparable to other prediction tools. This shows a positive opportunity for improvement in the near future.

3.5 CONCLUSION

In this study, machine-learning and computational intelligence approaches through HC and the ANN-Q (quantitative prediction) models were applied to predict the monthly WTI price for every barrel of crude oil in USD. The results obtained from the simulation study validate the effectiveness of the data selection process by the HC model. The HC model proposed a systematic approach that successfully extracts a comprehensive list of features used as the reference to the key factors that had caused the crude oil price market to be volatile. The quantitative prediction model's effectiveness and accuracy were derived from the utilisation of a good input variable combination as depicted in the ANN-Q model. The One-step Returns function, employed to reduce errors and noise, had proven to normalise the data and produced better prediction outcomes, compared to the usage of time-series data as a type of data representation. This chapter also had successfully proven the hypothesis that price data classification discussed in the HC model represents a pattern of rules that contributes to the market, as well as produces satisfactory prediction outcomes which reflect the good interconnections existing in the model. In addition, the results obtained from the prediction also proved the hypothesis regarding selecting single digit hidden neurons of 3, 4 and 5 in the network had helped to maintain good network propagation. The amalgamation of linguistic information with this quantitative prediction model in an extended model, discussed in section 5.5 of Chapter 5, will demonstrate interesting information for crude oil price market. Chapter 4 will introduce the derivation of linguistic elements from online news articles to be used as the linguistic inputs in the next prediction models.

Chapter 4 Content Utilisation and Sentiment Analysis of Data Mined from Google News in Fuzzy Grammar Fragments with the Development of the Fuzzy Expert System for the Rule-based Expert Model

Overview

The development of fuzzy grammar fragment extraction in this chapter is the heart of our linguistic prediction model. The fragments[115] *extracted from the news articles helped in the sentiment analysis phase of this chapter by exhibiting those activities or events surrounding the crude oil market that affect the pricing. This chapter discusses the utilisation of news content in deriving a new set of input data for training in the linguistic prediction model. This chapter aims to extract relevant features from news articles and then use them as input data for the linguistic prediction model by i) developing the fuzzy grammar for the fragment extraction and ii) developing the fuzzy expert system based on the sentiments of the extracted textual information.*

The results from this chapter are used to develop the linguistic prediction model in Chapter 5. In order to observe the effectiveness and the accuracy of the results obtained in the linguistic prediction model, the model was compared to results generated by the quantitative prediction model discussed in Chapter 3.

[115] Fragments: the meaning sequence (or vector) of features.

4.1 INTRODUCTION

Compared to stock market prediction, research concerning the price prediction of crude oil is limited. Nevertheless, the scarcity of the research is still of interest to researchers and practitioners. Future predictions are derived from historical information retrieved from past time-series data. Extracting the correct, precise and significant information is therefore an important challenge in the anticipation of future events, as is aggregating it into a decision system. Although there appears to be little research in the domain, others researchers have also implemented linguistic data as the input for their models. For example, Kroha, et al. [59] used text mining to extract information from financial business news for forecasting purposes, while Milea, et al. [60] used European Central Bank (ECB) statements to predict the MSCI EURO index[117] [61]. In addition, Schumaker and Chen [62] analyses textual information from financial news streams to predict the stock market. Through the investigation, these researchers believed that linguistic information such as news offers valuable insights into specific domain issues. They contained knowledge which is ready to be implemented in various kinds of intelligent ways. Due to the advances of technology, online news is cheap and accessible. This chapter will discuss the 'exploitation' process of such valuable knowledge in the models of this research domain.

For the purpose of this research, monthly articles from Google News were used as the linguistic input data to produce a prediction based on the sentiment of the fragments extracted. News items were mined for the keywords 'crude oil: price' from the chosen period of January 2007 to December 2010. Later in this chapter, we will explore the content and sentiment analysis and its utilisation in helping to forecast the following month's price.

[117] The leading benchmark for European equity funds [61].

4.2 CONTENT UTILISATION

Chapter 3 combined machine learning and computational intelligence approaches via the hierarchical conceptual (HC) and the artificial neural networks-quantitative (ANN-Q) models. The data were then used to predict the monthly West Texas Intermediate (WTI) price per barrel in US Dollars (USD). The results obtained from the research simulation validate the effectiveness of the data selection process in the HC model.

The HC model also successfully extracted a comprehensive list of key impact factors that have caused the crude oil price market to be volatile. The accuracy of the data selection process extracted factor vectors for the ANN-Q model to use in its learning and predicting module. Although promising results were obtained from the ANN-Q model, we hypothesise that linguistic factors play a vital role in contributing to price volatility. This chapter focuses on the analysis of the textual information contained in extended Google News articles obtained from the HC model. It is important for this model to gather precise and significant data in order for the prediction to be reliable and accurate. Simultaneously, this approach is designed to extract as much significant information as possible. In this extracted information lie the rules that govern the behaviour of the crude oil price market and which were later aggregated into a database in order to be utilised in the rule-based expert model. The rule-based expert model will be discussed in section 4.6 of this chapter.

The monthly linguistic information retrieved was also used to reveal the sentiment[119] of the articles. The analysed sentiment of an extracted fragment is useful for gaining insights into the underlying causes of price market volatility and is therefore important. These analysed sentiments are useful as input for anticipating the future direction of the crude oil price market. Thus, they were utilised as input data for the linguistic prediction model in Chapter 5.

The analysis began with the retrieval of textual information from Google News, the selection of appropriate categories from the General Inquirer (GI)[120] and the definition of words' compound definitions for building a terminal grammar[121],

[119] Sentiment: opinion or expression.
[120] General Inquirer (GI): a lexicon of word categories.
[121] Terminal grammar: a customised word dictionary.

followed by the application of the extracted fuzzy grammar fragments. Later in the process, it extracts rules from the fragments and finally identifies the sentiment of the fragments. All these processes are discussed in detail in subsections 4.2.1 and 4.4.2 of this chapter. The framework for the content utilisation and sentiment analysis processes is summarised in Figure 4.1, which exhibits the workflow of this analysis.

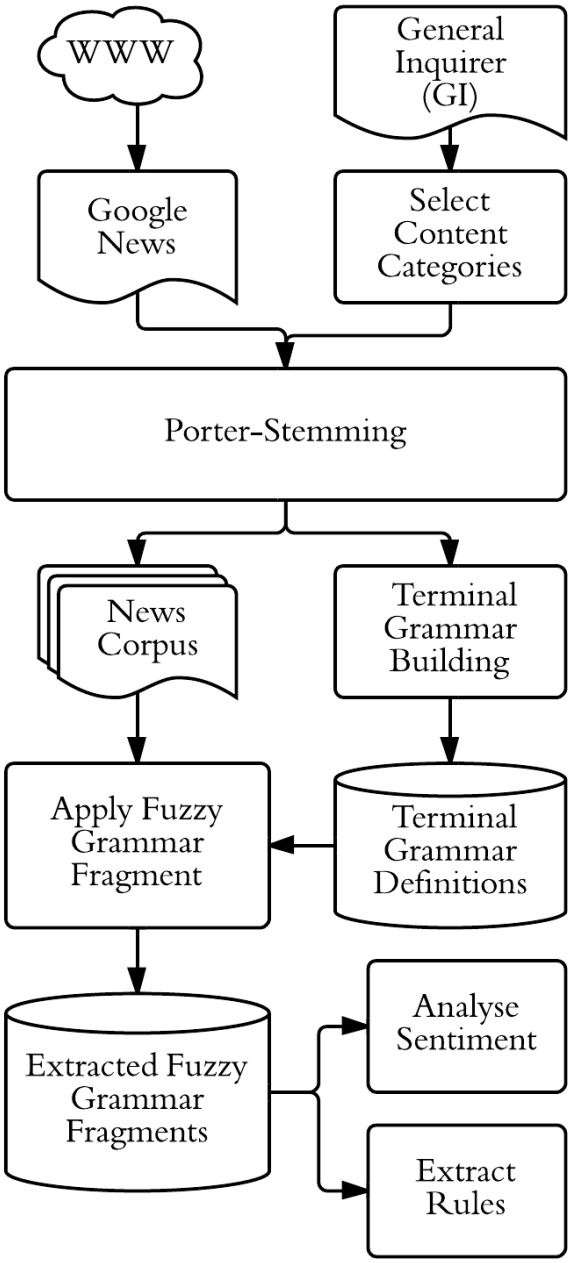


Figure 4.1 The Content Utilisation and Sentiment Analysis Framework for Google News.

4.2.1 Google News-Mining and News Corpus

"Google News is a computer-generated news site that aggregates headlines from more than 4,500 English-language news sources worldwide, groups similar stories together and displays them according to each reader's personalised interests" [63]

Google is well known as one of the biggest search engines in the world. It runs over one million data servers [64] that process more than one billion search requests every day [65] and is entitled to the latest news. Google.com (a US-focused site) was ranked the second-most often visited website in the World Wide Web by Alexa[125] in November 2012. Among its internet products and services is the Google News service. Google News is a corpus of all the computer-selected news and articles, which are evaluated according to the frequency of appearance on the online sites, and enables one to personalise search criteria according to one's preferences. As our research deals with time-series data, Google News is the perfect news miner because it has the ability to classify news according to specific periods and keywords.

The articles adopted in this chapter were arranged monthly for the period from 2004 to 2011 (8 years). The articles were extracted monthly, based on the following durations:

i) early in the month (days 1-10);
ii) the middle of the month (days 11-20); and,
iii) the end of the month (days 21-30/31).

An average of approximately five articles was extracted per time period, and 1,440 articles were collected from the Google News service for the durations mentioned in (i) to (iii) above, later classified into an annual corpus for each month.

After the analysis was made using the eight-year news corpus, this research decided to focus on the most recent four years (2007-2010), which contained 720 articles for the next process, namely the content and sentiment analysis. This period is believed to contain valuable information that explains the market's volatile behaviour and includes important historical events. In preparation for the next process, all the

[125] Alexa: a worldwide, online data traffic provider.

linguistic data collected will be pre-processed into the appropriate format. The process is discussed thoroughly in section 4.2.2.

4.2.2 Porter-stemming

A process that prepares the news corpus data and enables them to be utilised as the input data in this model was induced before the development of a rule-based expert model had begun. Porter-stemming[127], through its normalisation method, was used to reduce each word in the corpus to its root form. This normalisation process removes the most common morphological and inflexional endings that are contained in English words [66].

Normalising the words into their root form will help to simplify the process of identifying the depicted features in an article and will later simplify the process of extracting the grammar fragments from the articles. The process of normalising the word means distinguishing the English word's identity from its grammatical form. Each word in an English dictionary could be categorised into different categories, such as positive or negative.

The identification of the same word category after a Porter-stemming process is important, as it enables the extraction of fragments from the statement of an article. Before words were put into categories, the news articles that were mined and extracted were first Porter-stemmed. An example of a Porter-stemmed article is presented in Table 4.1, while the process of categorising the words in the article is introduced in the next section, subsection 4.2.3, as word category selection using the General Inquirer (GI).

[127] Porter-stemming: a process used to simplify English words into their root form. For example, the word "enlighten", after Porter-stemming, will retain only the section "light". The "en-" is the morphological commoner of the word "light", whilst "-en" is the inflexional ending.

Table 4.1 Example of a Porter-stemmed Article (January 2009).

ORIGINAL STATEMENT OF AN ARTICLE	PORTER-STEMMED STATEMENT OF AN ARTICLE
Crude Ends Up As Wall St Gains, Dollar Down. Sat Aug 1, 2009 1:16am IST NEW YORK, July 31 (Reuters) - U.S. crude futures gained more than 3 percent on Friday, moving up with Wall Street, after U.S. GDP data for the second quarter showed the economy contracted less than expected and bolstered hopes that the recession was easing.	Crude end up a Wall St gain dollar down Sat Aug 1 2009 1:16am IST NEW YORK Juli 31 (Reuters) US crude futur gain more than 3 percent on Fridai move up with Wall Street after US GDP data for the second quarter show the economi contract less than expect and bolster hope that the recess wa eas

4.2.3 General Inquirer (GI) Category Selection

In this chapter, the sentiment of a news fragment and its effect on the volatility of the price is observed. For this purpose, General Inquirer (GI) [67], which is a computer-assisted approach and a content analysis tool for textual information extraction, was employed.

The GI technique was first used in an economic context in [68]. The economic context was used to calculate the correlation between the focus on wealth and wealth-related words in the German Emperor's speeches between 1870 and 1914. It was later discovered that there was a strong frequency of occurrence of words related to wealth in the speeches delivered during that period. Eventually, GI was also used to satisfactorily visualise the evolution of eight dimensions of sentiment in [69]. The authors studied the concepts of Joy, Sadness, Trust, Disgust, Fear, Anger, Surprise and Anticipation as sentiments in web news and gained satisfactory results from the analysis.

The GI houses 11,794 lexical items and has over 300 content categories, derived from both the Harvard-IV Dictionary [70] and the Lasswell Value Dictionary [71] [72]. Of the 300 word categories provided by GI, only seven categories were chosen to be implemented in the content utilisation and sentiment analysis process. The seven categories chosen are listed in Table 4.2. These word categories were chosen to dictate the positive and negative sentiment of a fragment of an article. The correct dictation of a word category in the fragment helped to give an overview of the price market behaviour. Of the 300 word categories, positive and negative categories emerge as the most contained word category in the GI. Therefore, these two categories were chosen with the combination of other similar categories to simplify the process and make it more focused.

Table 4.2 Content Categories Selected from the General Inquirer (GI)

CATEGORY	SYMBOL	DEFINITION
Positiv	*p*	Consisting of 1,915 positive words
Negativ	*n*	Consisting of 2,291 negative words
Strong	*s*	Made up of 1,902 words implying strength
Weak	*w*	Made up of 755 words implying weakness
Increas	*i*	Containing 111 words indicating growth
Decreas	*d*	Containing 82 words indicating reduction
Activ	*a*	Enclosing of 2,045 words characterised by energetic work
Passiv	*ps*	Enclosing of 911 words characterised by visible action

Based on Table 4.2, despite the words being indexed according to their sentiment in the lexicon, it was discovered that the same lexical items in *s, i* and *a* were also shared by the *p* category; thus, similar words were also contained in different categories. Similar conditions were aligned to the *w, d* and *ps* categories, where the same lexical item was also shared by the *n* category. Thus, from the chosen categories, we discovered that the shared words could simply be reorganised into just two indices, *p* and *n* categories, as presented in Figure 4.2. We eliminated any repeated words from all the categories and reclassified them into two main content categories. This lexical reclassification resulted in 1,896 words in the *p* category and 2,039 words in the *n* category, which will be used as compounds[129] in the terminal grammar[130], as will be discussed in the next section of this chapter.

[129] Compounds: two or more words that are joined to form one word.
[130] Terminal grammar: a 'dictionary' that contains definitions and rules that are used to form a formal grammar.

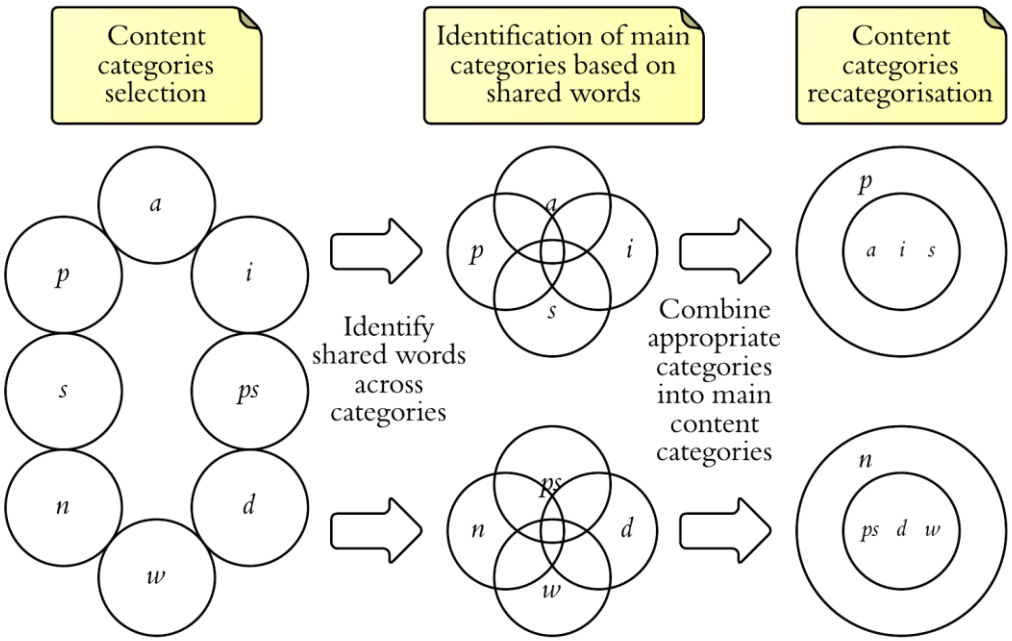


Figure 4.2 Categorisation Process of Lexical Items into Positive and Negative Content Categories.

4.3 TERMINAL GRAMMAR BUILDING

Terminal grammar is known to consist of a terminal (list of words) denoted by *t*, and nonterminal ('definition') symbols, denoted by *n*, as elements in the lexicon. These elements are employed in a module to produce rules for constituting a formal grammar. A formal grammar, or a context-free grammar, is a set of alphabet strings that give a precise description of a formal language. Since linguistic data are used as input in this chapter, constructing the grammar is essential for the model to be able to analyse and extract the sentiment. The essence of the research is to analyse fragments in an article and then assign the appropriate sentiment to them. Furthermore, it has been established that the length of a text fragment is shorter than is a sentence in an article, and that context-free grammar is good at handling small-sized linguistic data. Therefore, a context-free grammar was chosen and is engaged in understanding the linguistic data used.

4.3.1 Language Structure

A sentence in a language consists of structures that map the meaning of a statement. In English sentences, these are generally mapped according to the examples in Figure 4.3.

1 {THING *The oil demand*} {PLACE *in China*}
2 {THING *The oil price*} is {PROPERTY *volatile}*
3 {ACTION {THING *Oil demand*} *is increasing* {PLACE *in China*}}
4 {ACTION {THING *Oil price*} *went* {PATH *lower in the market*} {TIME *last month*}

Figure 4.3 An Example of English Language Structure

Figure 4.3 shows the general elements that construct the four different sentences. "THING", "PLACE", "PROPERTY", "ACTION", "PATH" and "TIME" are the structures that build the formal grammar of a sentence. In a general 'dictionary', these are the subjects of a sentence, whilst, "*The oil demand", "in China", "The oil price", "volatile", "lower in the market"* and "*last month"* are the predicates. A sentence is considered to be a sentence when there is a combination of both subject and predicate elements; grammatically, these are called nouns and verbs. The objective of fragment extraction is to extract a fragment with a specific grammar structure. This fragment extraction process is discussed thoroughly in section 4.4 of this chapter.

Words are often categorised according to traditional parts of speech, as tabulated in Table 4.3.

Table 4.3 Grammar Building Structure: Word Categories

CATEGORY	DEFINITION	EXAMPLE
Nouns	Names of things	Oil, price, demand, supply
Verb	Action or state	Become, hit, increase, decrease, maintain
Pronoun	Used in place of noun	I, you, we, them, they, those
Adverb	Modifies verb, adjective, adverb	Suddenly, very, accurately, drastically
Adjective	Modifies noun	Steady, volatile, instable
Conjunction	Joins things	And, but, while
Preposition	Relation to noun	To, from, into

These traditional categories are normally grouped into parts, and sometimes subpart of a sentence, called constituents. Constituents are usually phrases named according to the head of the constituent. Table 4.4 shows examples of constituent phrases. These constituent phrases are responsible for building the grammar of a language, which gives lingual[133] structure to a derived sentence.

Table 4.4 Grammar Building Structure: Examples of Constituent Phrases.

PHRASE	TYPE OF PHRASE	DEFINITION
The demand in China	Noun Phrase	The head, *demand,* is a noun
Extremely volatile	Adjective Phrase	The head, *volatile,* is an adjective
Reduce the production	Verb Phrase	The head, *reduce,* is a verb

The next sections discuss the application of context-free grammar (CFG) with regard to the crude oil price problem, together with the implementation of the language structure discussed in subsection 4.3.1.

[133] Lingual structure: the grammatical structure of a sentence. For example, a sentence is considered lingual if it follows the grammatical structure rule of having at least a noun and a verb in a sentence.

4.3.2 Context-Free Grammar (CFG)

The idea of basing a grammar on a constituent structure dates back to 1879, when a modern conceptualisation psychologist, Wilhelm Wundt [73], attempted to quantify human thoughts according to basic elements. He was inspired by success in the fields of chemistry and physics, in which constituent pieces such as atoms and elements were described. His dream of quantifying language was only formalised by Chomsky [74] [75] [76] and Backus [77] in 1956 and 1959, respectively. CFG is a well-known method of modelling constituency today, and is also known as Phrase Structure Grammar[134]. The Backus-Naur Form[135] is also based on CFG.

Referring to Figure 4.4, CFG structure is known to have the elements of

$G = <T, N, S, R>$

where *G, T, N, S* and *R* denote the grammar (generates a language, *L*), a terminal (a lexical set), a nonterminal, a start and rules/production with the form of $X \rightarrow \gamma$[136], respectively. In *R, X* constitutes a non-terminal, and γ, a sequence of terminals and non-terminals. Figure 4.4 shows examples of *T* elements in a *G* that are categorised as determiners, nouns, verbs and auxiliaries. The rules were regenerated to model the CFG in the crude oil price domain in Figure 4.5. The rules were then represented in the form of a parse tree, as depicted in Figure 4.6.

[134] Phrase Structure Grammar: a grammatical framework that defines the syntax and semantics of a language.
[135] Backus-Naur Form: a syntax description for languages used in computing.
[136] Classical IF <antecedent> THEN <consequent> rule.

1	*G*	=	*< T, N, S, R >*
2	*T*	=	{*the, a, is, when, price, supply, war, increases, plummets*}
3	*N*	=	{Sentence-S, Noun Phrase-NP, Nominative-NOM, Verb Phrase-VP, Determiner-Det, Noun-N, Verb-V, Auxiliary-Aux}
4	*S*	=	*S*
5	*R*	=	{
6	*S*	→	NP VP
7	S	→	Aux NP VP
8	S	→	VP
9	NP	→	Det NOM
10	NOM	→	Noun
11	NOM	→	Noun NOM
12	VP	→	Verb
13	VP	→	Verb NP
14	NOM	→	Noun
15	NOM	→	Noun NOM
16	VP	→	Verb
17	VP	→	Verb NP
18	Det	→	*the* | *a* | *when*
19	Noun	→	*price* | *supply* | *war*
20	Verb	→	*increases* | *plummets*
21	Aux	→	*is*
22			}

Figure 4.4 Example Of Context-Free Grammar (CFG).

Figure 4.4 shows the structure of a CFG that is designed to constitute a sentence constructed for oil price events. To understand the CFG, let *G* in (1) be a 'dictionary'. *G* is a 'dictionary' that consists of *T* words, which were defined by *N*, the definition. To construct a sentence *S*, words need to abide by grammatical rules as depicted by *R* in (5) and listing rules (6), (7) and (8). Rules (9) to (17) declassify the definitions made in *N* into smaller definitions and, finally, rules (18) to (21) dictate the definitions set in *N* with the words depicted in *T*. Based on the example in Figure 4.4, *S* could be generated by the following:

i) a noun phrase (NP) and a verb phrase (VP), or

ii) an auxiliary (AUX), a noun phrase (NP) *and* a verb phrase (VP), or

iii) a VP only.

An NP is normally constructed by a determiner (Det) and a nominative (NOM), while a VP is constructed by either (i) a verb or (ii) a verb and an NP. Meanwhile, the values of the Det, Noun, Verb and Aux are pre-defined, as in Figure 4.4.

Figure 4.5 shows an example of a sentence, *S*, with an NP and a VP as per the grammatical rule. In this example, the rule starts by addressing the 'definition' of an NP first; this consists of Det and NOM, and later continues with the definition of a VP. Every rule in the example represents the definition of each grammar that was pre-determined in Figure 4.4.

S → NP VP
→ Det NOM VP
→ *The* NOM VP
→ *The* Noun VP
→ *The price* VP
→ *The price* Verb NP
→ *The price increases* NP
→ *The price increases* Det NOM
→ *The price increases when* NOM
→ *The price increases when* Noun NOM
→ *The price increases when supply NOM*
→ *The price increases when supply plummets*

Figure 4.5 Examples of Context-Free Grammar (CFG) Rules For the Crude Oil Price.

The grammatical definitions were also presented as a hierarchy of rules in a parse tree, as illustrated in Figure 4.6. It delineates the derivation of grammars or rules for the sentence given in the example.

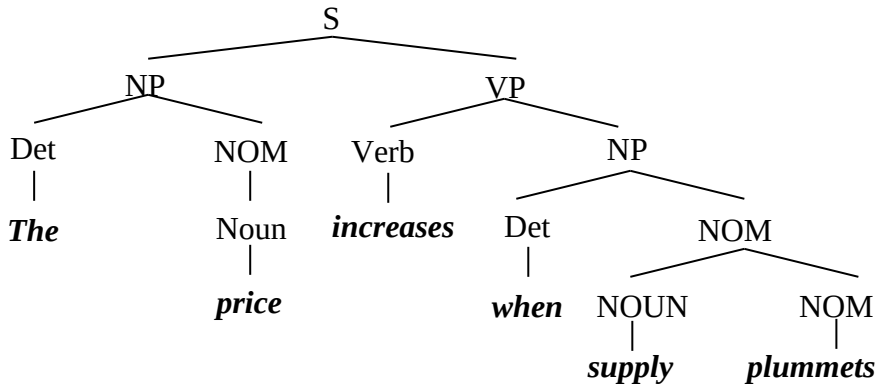


Figure 4.6 Example of a Context-Free Grammar (CFG) Parse Tree for the Crude Oil Price.

A sentence that is constructed by two CFGs defines a set of all the strings of words that can be derived from the grammar of a formal language. Therefore, a sentence is considered to be grammatical[140] if it produces word combinations as defined in the set and will be considered to be ungrammatical if the converse is the case. In this research, the sentiment of a fragmented news articles is measured based on the structural grammar of a text fragment.

These text fragments are based on a pre-defined terminal grammar. These text fragments are based on a pre-defined terminal grammar. They are normally shorter than the length of input used in [78].Hence, CFG is utilised in this research. CFG was chosen because of the ease of use when handling small arrays of words. In addition, it reduces the complexities of natural language processing, which simplifies the analysis process. This simplicity is imperative for organising the appropriate parsing process for the grammars to be used in the next stage of this research.

[140] Grammatical: abiding by the rules of grammar.

4.3.3 Grammar Parsing

In order to understand the structure of a grammar, grammar parsing is useful to expose the framework by which a sentence is constructed. Parsing is the process of searching and analysing the structure of a grammar. Searching the structure begins by rewriting the grammar (rules) in order to find structures matching the input string of words encapsulated in the domain sentence. Parsing can be done in two ways, namely top-down or bottom-up. Table 4.5 shows an example of bottom-up parsing using the shift-reduce[141] method with the crude oil price as the domain. This shift-reduce method is similar to the process in the CFG illustrated by Figure 4.5 and Figure 4.6, except that the process was conducted the other way around.

In the CFG, the structure of a sentence was first defined before assigning each word a grammatical definition as defined in the terminal grammar[142]. A grammar parsing using the bottom-up method, on the other hand, utilises a sentence to search for the structure[143]. Grammar parsing is considered to be successful whenever it ends with the symbol Start (S) in the stack[144]. The symbol S at the end of the stack means that the searching of a grammatical structure from the bottom to the top is complete, and all the words in the sentence have been designated with definitions.

[141] Shift-reduce: a table-driven, bottom-up parsing method.
[142] Constructing a structure to find or extract a sentence.
[143] To use a sentence to extract structure.
[144] Stack: containing the grammatical symbols or rules.

Table 4.5 Grammar Parsing: Examples of Bottom-Up, Shift-Reduce Parsing.

	ACTION	STACK	INPUT REMAINING
1	Shift	( )	*Oil price increases when supply plummets*
2	Reduce, Noun→*oil price*	(*Oil price*)	*increases when supply plummets*
3	Shift	(Noun)	*increases when supply plummets*
4	Reduce, Verb→*increase*	(Noun *increases)*	*when supply plummets*
5	Shift	(Noun Verb)	*when supply plummets*
6	Reduce, Det→*when*	(Noun Verb *when*)	*supply plummets*
7	Shift	(Noun Verb Det)	*supply plummets*
8	Reduce, Noun→*supply*	(Noun Verb Det *Supply*)	*plummets*
9	Shift	(Noun Verb Det Noun)	*plummets*
10	Reduce, NOM→*plummets*	(Noun Verb Det Noun *plummets*)	-
11	Reduce, NOM→Noun NOM	(Noun Verb Det Noun NOM)	-
12	Reduce, NP→Det NOM	(Noun Verb Det NOM)	-
13	Reduce, VP→Verb NP	(Noun Verb NP)	-
14	Reduce, S→VP	(Noun VP)	-
15	Reduce, NOM→Noun	(Noun)	-
16	Reduce, NP→NOM	(NOM)	-
17	Reduce, S→NP	(NP)	-
18	SUCCESS	(S)	-

To apply the CFG or bottom-up method, a terminal grammar that consists of specific grammar definitions was built to comply with the specifications related to the domain. It is important to define grammatical definitions precisely in the domain's terminal grammar, so as to ensure that appropriate grammars were parsed to extract the correct text fragment from a sentence in an article. The extracted fragments derived from this process were later used as input in the sentiment analysis.

The process is designed to do the following:

i) to disregard any insignificant words, as well as to focus on and discard noise;
ii) to certify that only appropriate text fragments extracted. These will contain only the core terms that correspond to the key events related to the crude oil market; and
iii) to quantify the directions of a fragment so as to explore the essence of its sentiment.

Hence, our terminal grammar includes:

i) any words—a regular expression defines the non-numeric strings;
ii) content categories—a class that defines the sentiment of a fragment; and
iii) core terms—a lexicon that defines the terms and events that significantly contribute to the volatility of the price market.

These points will be discussed comprehensively in section 4.3.4 of this chapter.

4.3.4 Terminal Grammar Definitions

Generating simple grammars from simple sets of data is required in order to build the grammar. For this purpose, a degree of knowledge was applied to identify the common word sets that belong to its pre-defined definition [79]. Our aim was to simplify the complexity of a grammatical definition so as to emphasise only the attributes that could achieve the objective of analysing the sentiment of a fragment and quantifying its direction. Hence, the focus is on three main elements, as follows:

i) any words—a regular expression that defines any non-numeric strings;

ii) content categories—a class that defines the sentiment direction of a fragment; and

iii) core terms—a lexicon defines the terms and events that contribute significantly to the volatility of the price market based on the hierarchical conceptual (HC) model in section 3.2 of Chapter 3.

The basic mechanism(s) inherit in the process are to disregard any insignificant words in the article and to focus on and discard noise, to certify that only appropriate text fragments, which contain the core terms corresponding to the key events of the domain, are extracted and to quantify the direction of a fragment in order to explore the essence of its sentiment. Subsequently, this process will result in a grammatical derivation that tags the atomic expressions as a parsed word in a fragment that contains an appropriate definition derived from the terminal grammar. To compensate for the undefined words in the terminal grammar, an atomic expression was also derived for these alphabetical strings. All other words that are not specifically defined will be denoted as ‘any word’, or *aw* in the grammar.

Table 4.6 shows the content definitions of the terminal grammar used in the research. This terminal grammar was later compiled into a document named *header*. *Header* is called and accessed by the extraction module to parse the grammar in an article and later to extract fragments of text from a sentence. The example of the *header* file is annexed in the Appendix A.

Table 4.6 The Content Definitions of a Terminal Grammar.

TYPE	DEFINITION
Content Categories:	
Positive	Combination of *positive* | *strong*| *increase* and *active* categories chosen from GI. Consisting of 1,915 positive words
Negative	Combination of *negative* | *weak* | *decrease* and *passive* categories chosen from GI. Consisting of 2,291 negative words
Core terms:	
oilFactor	Factors contributing to the market fluctuations. *Crudeoil* | *oilprice* | *crudeoilprice* | *consumption* | *demand* | *import* | *inventory* | *oecd* | *opec* | *population* | *production* | *refinery* | *reserves* | *stock* | *supply* | *war* | *weather* | *recession*
economicTerm	Factors contributing to economy *Economy* | *recession* | *forex* | *currency* | *downturn* | *inflation* | *gdp* | *gbp* | *euro* | *usd* | *usdollar* | *cpi* | *speculation*
weatherNaturalDisasterTerm	Terms related to natural disaster and weather *Weather* | *spring* | *summer* | *autumn* | *winter* | *hot* | *cold* | *freeze* | *storm* | *frost* | *hurricane* | *flood* | *fire* | *rain*| *temperature* | *warm* | *mild* | *fog* | *tsunami* | *earthquake* | *tornado* | *heat wave* | *eruption* | *avalanche* | *gale* | *twister* | *Typhoon* | *windy*
opecCountries	Members of The Organisation of Petroleum Exporting countries (OPEC) *Algeria* | *Angola* | *Ecuador* | *Iran* | *Iraq* | *Kuwait* | *Libya* | *Nigeria* | *Qatar* | *Saudi Arabia* | *United Arab Emirates (UAE)*| *Venezuela*
importingCountries	Main petroleum importing countries *China* | *India* | *United States (US)* | *Europe* | *Japan* | *Germany* | *Spain* | *France* | *Italy* | *Taiwan*
politicalTerm	Terms related to politics *Geopolitic* | *politic* |*violence* | *war* | *strike* | *air strike* | *terror* | *terrorist* | *riot* | *military* | *militant* | *army* |*bombing* | *revolution* | *attack* |*crisis* | *picket* | *conspiracy* | *protest* | *elections* | *conflict*
aw	Any other words in the article Derived from a regular expression of [*a-zA-Z]+\D*, where a-z* denotes words with lower capital alphabets and *A-Z* denotes the opposite. *\D* is used to parse any alphabets other than 0-9 with + and * as the repetition parsing

It is vital for the terminal grammar to be appropriately defined. In order to extract the relevant fragments from a sentence correctly and to be utilised as input in the sentiment analysis process, an appropriate grammar needs to be parsed. An inappropriately defined grammar leads to noise and divergence in the fragment extraction process, which later disrupts the process of obtaining accurate results. Section 4.4 discusses the development aspect of this fragment extraction module further.

4.4 THE APPLICATION OF FUZZY GRAMMAR FRAGMENT EXTRACTION FOR THE FUZZY EXPERT MODEL

An appropriate terminal grammar definition is essential in our sentiment analysis process. An appropriate definition leads to appropriate grammar parsing, which later helps to extract an appropriate grammar fragment for analysis. This section discusses in detail the development of our customised grammar fragment extraction module. This module is an adaptation of a fuzzy grammar fragment extraction module developed by [80], where the algorithm developed by the author was employed to suit the crude oil market domain. This section discusses the development of the text fragment identification framework, while the grammar fragment extraction process is presented in Figure 4.7. The framework presents the overall structure for deriving the fuzzy expert model with the application of fuzzy grammar fragment extraction, which is discussed further in section 4.7 of this chapter. There were four structures (denoted by the numbers 1 to 4 in Figure 4.7) involved in developing the fuzzy expert model:

1) text fragment extraction process[145];
2) grammar fragment extraction process[146];
3) sentiment mining and analysis process[147]; and
4) the rule extraction and delineation process[148].

The text and grammar fragment extraction framework shown in Figure 4.7 has the following aims:

i) to learn the grammar fragments structure- by identifying the constituents based on the grammatical definitions defined in the terminal grammar; and
ii) to then constitute the grammar into rules for extraction, in order to analyse its sentiments and to later quantify its directions.

[145] Discussed in subsection 4.4.1.
[146] Discussed in subsection 4.4.2.
[147] Discussed in section 4.5.
[148] Discussed in section 4.6.

Achieving these aims involves both the learning phase and the testing phase of the fragments. Detailed explanation about Figure 4.7 is discussed in subsections 4.4.1 to 4.6.

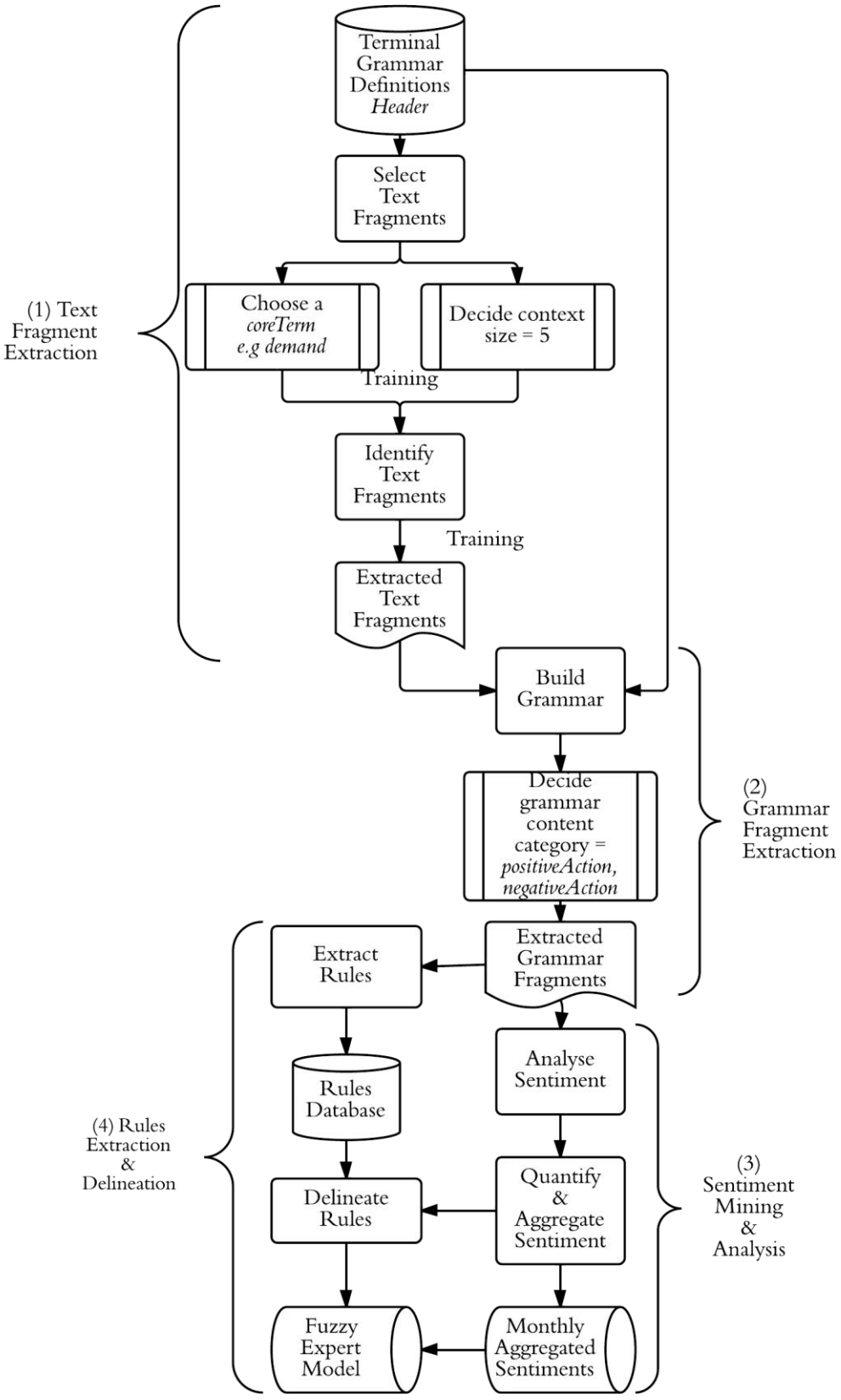


Figure 4.7 The Fuzzy Expert Model Framework with Text Fragment and Grammar Extraction, Sentiment Mining and Analysis and Rule Extraction Structures.

4.4.1 Text Fragment Selection and Extraction

News articles require some degree of preparation before they are ready for fragment extraction. Referring to structure (1) in Figure 4.7, in order to extract appropriate fragments, the correct conditions are ensured to exist for the extraction to be significant and relevant. Therefore, a fragment of a sentence in an article was selected based on two conditions:

i) it must contain one or more *coreTerm*[149]*;* and

ii) the size of the context must equal to five[150].

The aim of the selection training was for the module to identify the depicted conditions in a sentence in order to extract relevant fragments from an article.

Another aim of this research was to discover related events associated with the crude oil market. The *coreTerm* defined by *oilFactor* in the terminal grammar and tabulated in Table 4.7 was chosen as the header[151] when training the fragments. The context size, or the length of the fragment to be extracted before or after a header, was limited to only five words. This means that the text fragments are automatically selected from the news article based on a maximum of five words preceding a *coreTerm*, five words succeeding a *coreTerm* or five words preceding and succeeding a *coreTerm*. An example of this extraction based on the chosen context size and *coreTerm* is presented in Table 4.7. The text fragment extraction process discussed in this section and the grammar fragment extraction process discussed later in section 4.4.2 were done with the application of the fragment-extraction algorithm developed by [80].

[149] *coreTerm = oilFactor:* the 'noun' or the 'terminal' element defined in the terminal grammar, tabulated in Table 4.5.
[150] The context of the fragment must only include five words before a *coreTerm*, five words after a *coreTerm*, or five words before and after a *coreTerm*, as tabulated in Table 4.6.
[151] Header: an identity that needs to be parsed in the grammatical structure of a sentence in order to extract a fragment.

Table 4.7 Example of Extracted Text Fragments Derived from the Training with Content Size= 5, *coreTerm= OilFactor= Demand* for a News Article extracted in January 2008.

TYPE	STRUCTURE	EXTRACTED FRAGMENT
5 predecessors	*predecessor- predecessor- predecessor- predecessor- predecessor-**oilFactor***	of a recession hammering oil ***demand***
5 successors	***oilFactor** -successor- successor- successor successor- successor*	***demand*** for gasoline fell last week
5 predecessors and 5 successors	*predecessor- predecessor- predecessor-predecessor- predecessor- –**oilFactor**- successor- successor- successor- successor- successor*	and many observers expecting further ***demand*** deteriorate as the economy slow

To retain the meaning of a sentence, fragments were extracted based only on words that were included in the same sentence. Section 4.4.2 discusses the investigation process of a grammar-fragment extraction of the chosen monthly articles.

4.4.2 Grammar Fragments Building and Extraction

To be able to parse a grammar, a terminal grammar needs to be pre-defined appropriately. By referring to Figure 4.8, in order to be able to parse a grammar from a fragment, a *header* is called into the algorithm to build the grammar first. At this level, grammar was parsed and derived from the extracted text fragments according to the definitions set out in the terminal. Any other words that were not defined in the terminal were given a trailing <*aw*> to denote 'any other words'.

In order for the text fragment to derive a grammar fragment, training with the aim of extracting the grammar combination of a *coreTerm* and a *contentCategory*[152] from a sentence was set. This pattern aimed to extract and then analyse the events that contributed to the market's volatility from the news articles, with *oilFactor* chosen as the header in order to sense its sentiment and to quantify its direction, and a *contentCategory* was chosen for the sentiment of the event(s) described in the article.

A grammar derivation derived a grammar fragment that enabled the module to annotate the fragment with either a positive or a negative sentiment. The *positiveTerm* and *negativeTerm* that were defined in the terminal grammar as the *contentCategory* were used to pattern-match the grammar combination. A grammar fragment was extracted whenever a pattern was recognised as a grammar combination. All the <*aw*> trailing, denoting any other words in the fragment, were excluded from the grammar fragment extraction so as to reduce noise and to simplify the process. The algorithm used for both the text and for the grammar fragment extraction process is presented in Figure 4.8.

[152] *contentCategory*: the 'verb' or the 'non-terminal' element of a fragment that was defined in the terminal grammar.

```
1  Initialise   s_x as new input text
2  Initialise   threshold as input for fragment extraction
3  Initialise   input TG={tg_1,tg_2,...,tg_x,...,tg_a}as a set of learned grammar
4  Initialise   Results as the set of extracted fragments that have members
                of the learned grammars above the threshold value
5  Let          W={w_1,w_2,...,w_i,...,w_m} a finite set of words in s_x where
                Wm= splitString(s_x)

6                FOR each x-th grammar in TG
7                      Initialise min as the minimum length of the target grammar,
                              max as the maximum length of the target grammar
8                      Initialise start= 0
9                      Initialise inputEnd= number of words in s_x
10                     Initialise window= 5 * max
11                     WHILE (start<= inputEnd - min)
12                           IF (start+window-1>=inputEnd)
13                              THEN window=inputEnd-start
14                           END IF
15                           Initialise end= start+window-1
16                           Initialise batch [window]
17                                 FOR (a = start; a<=end; a++)
18                                       batch[a-start]= w[a]
19                                 END FOR
20                           Results.add(calcMaxMembership(batch,tg_x,threshold)
21                           start=start+max
22                     END WHILE
23               END FOR
```

Figure 4.8 The Algorithm used for Fuzzy Text Fragment Extraction Developed by [80].

The algorithm laid out in Figure 4.8 contained two inputs, namely s_x, the string of words, and the value of similarity threshold.

In steps (1) to (7), the algorithm started by highlighting and extracting only the relevant fragments that corresponded to the learned grammar. Next, a variable initialised as *Results* was set as the corpus for the extracted fragment. For this research, the initialisation was based on the *oilFactor* vectors defined in the terminal grammar, which were trained based on a specific month of the year, such as *Demand* for January 2008. Later, the sentences were split into W_m*;* an array of words in a fragment and the minimum and the maximum length of a fragment for every grammar learned were depicted. Steps (8)-(10) initialised three additional variables:

i) the 'start' marks the starting position of the window parser;
ii) inputEnd marks the end of the fragment identification by the window parser; and
iii) 'window' sets the maximum length of the window size to 5*max.

Later, in steps (11)-19), the sequence of the text fragments was prepared and, finally, the fragments with the highest membership threshold were filtered, highlighted and extracted in steps (20) and (21). Table 4.8, Table 4.9 and Table 4.10 show examples of the results obtained from the grammar fragment extraction process. The sentiments and rules, which were later utilised as input in the next model discussed in Chapter 5, were extracted along with these grammar fragments

Table 4.8 Examples of Extracted Grammar Fragments

with Content Size= 5, *coreTerm= oilFactor*= Supply, *contentCategory = positiveTerm*, *negativeTerm*.

DATE	tF	psetF	eGF	EDITED eGF	SENTIMENT	EVENT
2008 18/12	steps it takes to **reduce supply** to shore up prices risks	*step it take* ***reduc suppli*** *shore up price risk*	*aw-aw-aw-* ***negativeTerm-oilFactor****-aw-positiveTerm-oilFactor-aw*	***negativeTerm-oilFactor****-positiveTerm - oilFactor*	negative (0) rule= reduce supply→price up	to shore up prices risk
22/12	from Saudi Arabia that OPEC **supply cuts** will stabilise the price	*from saudi arabia that opec* ***suppli cut*** *will stabilis the price*	*aw-opecCountries_-aw-oilFactor-* ***oilFactor-NegativeTerm****-aw-PositiveTerm- aw oilFactor*	*opecCountries_-oilFactor-* ***oilFactor-NegativeTerm****-PositiveTerm-oilFactor*	negative (0) rule= supply cut →price stabil	OPEC cut supply
	But **supply cutbacks** have yet to stem	*but* ***suppli cutback*** *have yet stem*	*aw-* ***oilFactor - NegativeTerm****- aw-aw- aw-aw*	***oilFactor - NegativeTerm***	negative (0) rule= supply cutbacks	-
29/12	day in New York on **supply concerns** after Israeli air strikes	*dai in new york on* ***suppli concern*** *after israeli air strike*	***oilFactor-NegativeTerm****-aw-aw-aw-politicalTerm*	***oilFactor-NegativeTerm****-politicalTerm*	negative (0) rule= Israel air strikes→ supply concern	Israeli air strikes
2009	violence in oil	*violenc in oil*	*politicalTerm-*	*politicalTerm-*	negative (0)	violence in oil

5/1	exporter Nigeria and **supply problems** in the key US	*export nigeria and **suppli problem** in the kei us*	*oilFactor -aw-opecCountries-aw-**oilFactor**-**NegativeTerm**-aw-aw-aw-importCountries*	*oilFactor -opecCountries-**oilFactor**-**NegativeTerm**-importCountries*	rule= violence in Nigeria→ supply problems	exporter Nigeria
	OPEC decided to **cut supply** by 2 million barrels	*opec decid **cut suppli** by 2 million barrel per*	*oilFactor-aw-**NegativeTerm**-**oilFactor**-aw-aw*	*oilFactor-**NegativeTerm**-**oilFactor**-*	negative (0) rule= cut supply	OPEC cut supply by 2 million barrels

***tF** = text fragment, **psetF** = porter stemmed extracted text fragment, **eGF** = extracted grammar fragment.

Table 4.9 Examples of Extracted Grammar Fragments

with Content Size=5, *coreTerm= oilFactor*= War, *contentCategory = positiveTerm, negativeTerm.*

DATE	tF	psetF	eGF	EDITED eGF	SENTIMENT	EVENT
2008 7/6	rate climbs and prospects of **war** in Middle East **grow**, sparking	*rate climb and prospect of* ***war*** *in middl east* ***grow*** *spark*	*PositiveTerm-aw-aw-aw-* ***oilFactor****-aw-aw-aw-aw-* ***PositiveTerm***	*PositiveTerm- -* ***oilFactor-PositiveTerm***	negative (0) rule= war grow	Prospect of war in Middle East
5/7	opec **warned war** with iran would cause 'unlimited'	*opec* ***warn war*** *with iran would caus 'unlimited'*	*oilFactor_-* ***negativeTerm-oilFactor****-aw-opecCountries*	*oilFactor-* ***negativeTerm oilFactor****-opecCountries*	negative (0) rule= warned Iran war	OPEC warned war with Iran
2009 4/1	**Fears** about **war** in the Middle East, refinery cutbacks	***fear*** *about* ***war*** *in the middl east refineri*	***NegativeTerm****-aw-****oilFactor****-aw-aw-aw-aw-aw-oilFactor-negativeTerm*	***NegativeTerm****-aw-****oilFactor****-aw-aw-aw-aw-aw-oilFactor-negativeTerm*	negative (0) rule= fear war → refinery cutbacks	Fear of war in the Middle East

***tF =** text fragment, **psetF =** porter stemmed extracted text fragment, **eGF =** extracted grammar fragment.

Table 4.10 Examples of Extracted Grammar Fragments

With Content Size= 5, *coreTerm= oilFactor=* Price, *contentCategory= positiveTerm, negativeTerm.*

DATE	tF	psetF	eGF	EDITED eGF	SENTIMENT	EVENT
2009 5/11	stocks drop as **crude oil price plunges**	*stock drop* ***crude oil price plung***	*oilFactor-negativeTerm-* ***oilFactor-negativeTerm***	*oilFactor-negativeTerm-* ***oilFactor-negativeTerm***	negative (0) rule= stocks drop → crude oil price plunges	Stocks price dropping
9/11	Gulf storm also pushed the **price higher**	*gulf storm also push the* ***price higher***	*aw-weatherNaturalDisaster-aw-negativeTerm-aw-* ***oilFactor-positiveTerm***	*aw-weatherNaturalDisaster-aw-negativeTerm-aw-* ***oilFactor-positiveTerm***	positive (1) rule= gulf storm → price higher	Storm in Gulf
18/11	**Oil prices rise** on inventory concerns	***oil price rise*** *on inventori concern*	***oilFactor-positiveTerm-*** *aw-oilFactor-negativeTerm*	***oilFactor-positiveTerm-*** *oilFactor-negativeTerm*	positive (1) rule= inventory concerns →oil prices rise	Inventory concerns
24/11	**Oil prices advance** on weaker US dollar	***oil price advanc*** *on weaker us dollar*	***oilFactor-positiveTerm-*** *aw-negativeTerm-importCountries-economicTerm*	***oilFactor-positiveTerm-*** *negativeTerm-importCountries-economicTerm*	positive (1) rule= weak US Dollar → oil prices advance	Weak US Dollar
3/12	**crude oil price slide**	***crude oil price slide***	***oilFactor-negativeTerm-****aw-*	***oilFactor-negativeTerm-***	negative (0) rule= inventory	Inventory gains

	on inventory gains	*on inventori gain*	*oilFactor-PositiveTerm*	*oilFactor-PositiveTerm*	gains → crude oil price slide	
18/12	**Oil price falls** as USD jumps	***oil price fall*** *usd jump*	***oilFactor-NegativeTerm-*** *economicTerm-positiveTerm*	***oilFactor-NegativeTerm-*** *economicTerm-positiveTerm*	negative (0) rule= USD jumps → oil price fall	US Dollar increase
24/12	on BSE, after **crude oil prices surged** more than 3%	*on bse after* ***crude oil price surg*** *more than 3% on*	*aw-aw-aw-****oilFactor-positiveTerm****--aw-aw-aw*	***oilFactor-positiveTerm***	positive (1) rule= crude oil prices surged	-

***tF =** text fragment, **psetF =** porter stemmed extracted text fragment, **eGF =** extracted grammar fragment.

4.5 SENTIMENT MINING AND ANALYSIS WITH RULES EXTRACTION

The aim of sentiment analysis is to extract the sentiment from a fragment of a statement in an article. This analysis is important because it reveals the overall sentiment of monthly news articles in terms of a positive or negative sense. The analysed sentiments from the extracted fragments helped to delineate rules for the rule-based expert model tailored to the crude oil price domain. Andreevskaia and Bergler in [81] were able to assign the fuzzy categories of Positive and Negative to a set of words using the Sentiment Tag Extraction Program (STEP) based on extracted fuzzy sentiments. In addition, the approach in [82] starts with the extraction of a term's subjectivity and orientation from text, which extends the training sets that consist of positive and negative words with WordNet. They used a supervised, binary-classified learner to allocate the vector categories into Positive and Negative categories. Meanwhile, Sakaji, et al. [83] developed an automated extraction for expressions concerning economic trends. The extraction was a success, as it was able to categorise positive and negative expressions linked to the economic trends without a dictionary.

In this research, we employed a semi-automated sentiment analysis of a fragment, based on the grammatical rules extracted. Earlier, we predefined our terminal grammar to identify lexical items that define positive and negative sentiments in a fragment. These definitions were helpful for grammar parsing, which later derived a grammatical definition for each lexical item contained in the extracted fragment. The text fragments were trained to extract every grammatical combination of a *coreTerm (oilFactor*) selected for training with a *contentCategory* (*positiveTerm* or *negativeTerm*) according to the definitions in the terminal grammar. The sentiment of a fragment is determined based on grammar that defines a word according to either of the *content categories* in the fragment. Nevertheless, if more than one *contentCategory* was derived in the grammar fragment, the sentiment of the fragment is determined based on the 'verb' (*contentCategory)* of the 'noun' (*coreTerm*). Afterwards, the analysed sentiments were quantified, as specified in equations (4.1) and (4.2) below, to then be collected into a database for the rule delineation process. These quantified sentiments were also utilised as the input for the linguistic prediction model discussed in Chapter 5.

$$Sentiment_{neg} = f(coreTerm_{trained})(negativeTerm) = 0; \quad (4.1)$$

$$Sentiment_{pos} = f(coreTerm_{trained})(positiveTerm) = 1 \quad (4.2)$$

Every extracted fragment that contained $coreTerm_{trained}$ attached to a *negativeTerm* denotes a sentiment, 0 (negative value), and the opposing sentiment, 1 (positive value), if it is attached to a *positiveTerm.* Tables 4.8 to 4.10 in subsection 4.4.2 present examples of the grammar fragments that were extracted based on their selected *coreTerm*s, together with their assigned sentiment and extracted grammar rules. Moreover, this sentiment analysis process also successfully extracted useful events and rules concerning market volatility. In the next section, the knowledge gained from the domain is presented to an expert model in order to derive a form of application that is more schematic.

4.6 RULE DELINEATION

Knowledge is the theoretical or practical understanding of a subject or a domain. In order to represent that knowledge, one needs to formulate it into a basic explanatory rule. In the previous sections in this chapter, we gained knowledge regarding the domain market that can be structured into a fuzzy expert model. The analysed sentiments, together with the extracted rules from the grammar fragments that were transferred to a database, are still basic. Hence, this section discovers the process of representing the expert knowledge in a more useful and systematic form so as to prepare it for the development of our fuzzy expert model.

4.6.1 Rule Delineation with the Decision Tree

Delineating rules is an important process for interpreting the knowledge extrapolated from the grammar fragments into a systematic rule database. A decision tree, which is a transparent method of classifying observations, is employed to map this reasoning process. This systemic rule delineation will help to map the routes, leading to a better understanding of the events that were involved in contributing to the volatility of the crude oil price market.

Table 4.11 Example of a Database for Analysed Sentiments Derived from Extracted Grammar Fragments with *coreTerm = oilFactor*, October 2009

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	CORETERM	CONSUMPTION	DEMAND	IMPORT	INVENTORY	OECD	OPEC	POPULATION	RODUCTIO	REFINERY	STOCK	SUPPLY	WAR	WEATHER	RECESSASION	PRICE
	PORTER-STEMMED CORETERM	CONSUMPT	DEMAND	IMPORT	INVENTORI	OECD	OPEC	POPUL	PRODUCT	REFINERI	STOCK	SUPPLI	WAR	WEATHER	RECESS	PRICE
1	SENTIMENT	1	1	1	1		1		1	0	1	0			1	1
2	SENTIMENT		1		0				0	0	0	0			1	1
3	SENTIMENT		1		0				1	0		1				1
4	SENTIMENT		1		0											1
5	SENTIMENT		1													1
6	SENTIMENT															1
7	SENTIMENT															1
8	SENTIMENT															1
9	SENTIMENT															1
10	SENTIMENT															1
11	SENTIMENT															1
12	SENTIMENT															1
13	SENTIMENT															1
14	SENTIMENT															1
15	SENTIMENT															1
16	SENTIMENT															1
17	SENTIMENT															1
TOTAL (+) FRAGMENTS		1	5	1	1	0	1	0	2	0	1	1	0	0	2	17
TOTAL (-) FRAGMENTS		0	0	0	3	0	0	0	1	3	1	2	0	0	0	0
TOTAL FRAGMENTS		1	5	1	4	0	1	0	3	3	2	3	0	0	2	17
(+) = 1, (-) = -1,		Positive	Positive	Positive	Negative	Neutral	Positive	Neutral	Positive	Negative	Neutral	Negative	Neutral	Neutral	Positive	Positive
(+/-) = 0		1	1	1	-1	0	1	0	1	-1	0	-1	0	0	1	1
(+) CERTAINTY		1.0000	1.0000	1.0000	0.2500	0.0000	1.0000	0.0000	0.6667	0.0000	0.5000	0.3333	0.0000	0.0000	1.0000	1.0000
(-) CERTAINTY		0.0000	0.0000	0.0000	0.7500	0.0000	0.0000	0.0000	0.3333	1.0000	0.5000	0.6667	0.0000	0.0000	0.0000	0.0000

Table 4.11 is an example of a database for the analysed sentiments composed from the process in section 4.5. The sentiments for each *coreTerm* were accumulated according to the extracted grammar fragments in the grammar fragment extraction module. Based on Table 4.11, the membership functions for the sentiments of each *coreTerm* were also obtained from the analysis. The membership functions derived from the sentiment aggregation were: i) *negative;* ii) *neutral;* and iii) *positive*, which were quantified as {-1, 0, 1}, respectively. These membership functions and the fuzzy sets highlighted in Table 4.11 in pink and purple, respectively, are useful in that they can be exploited in the fuzzy inference system in order to build the fuzzy expert model.

The calculation for the aggregated sentiments was made to signal the overall sentiment of the selected *coreTerm* for that particular month. Nonetheless, the sentiment accumulation values for these fragments were quantified by equation (4.3).

$$S^{sign} = \begin{cases} 1, & \text{if } F^{pos} \geq F^{neg} \\ 0, & \text{if } F^{pos} = F^{neg} \\ -1, & \text{if } F^{pos} \leq F^{neg} \end{cases}$$

where,

F^{pos} as the positive fragment of a $coreTerm$;

F^{neg} as the negative fragment of a $coreTerm$. (4.3)

For the rule delineation process, 375 extracted fragments, which consist of 4,875 observations from the time range of January 2008 to December 2009, were used. The *coreTerms* used for this process, which were also used in sections 4.4 and 4.5, are displayed in Table 4.12.

Table 4.12 The Inputs = *coreTerm* = *oilFactor* Data used for Decision Tree with *Price* as the Decision Output

NO.	VARIABLE	INPUT [*OILFACTOR*]
	D^{T153}	**TOTAL DEMAND**
1	d_1	Demand
2	d_2	Consumption
	I^T	**TOTAL INVENTORY**
3	i_1	Inventory
4	i_2	Imports
	S^T	**TOTAL SUPPLY**
5	s_1	Supply
6	s_2	OPEC decisions on productions
7	s_3	Production
8	s_4	Refinery
9	s_5	Stocks
	P^T	**TOTAL POLITICS**
10	p_1	War
	W^T	**TOTAL WEATHER**
11	w_1	Weather
	E^T	**ECONOMY**
12	e_1	Recession
13	e_2	US Dollar
14		Population
15		OECD consumptions

Of the 15 *coreTerm*s listed for training, only 13 *oilFactor* features were used for this purpose as input, while *Price* was the decision output. The remaining two *oilFactor* features, namely OECD[154] and population, were omitted from this process because no fragments that corresponded to the defined grammar rules were extracted from the news articles, which indicates that there was no information derived from these two *oilFactors.* To analyse the sensitivity of the data used in this model, a series of simulation tests were conducted, thus testing the training: testing percentage data ratios of 90:10 per cent (90 per cent of total data were used for training and 10 per cent for testing), 80:20 per cent (80 per cent of total data were used for training and

[153] T is the accumulated value of each *oilFactor* feature.
[154] OECD: Organisation of Economic Co-operation and Development.

20 per cent for testing) and 70:30 per cent (70 per cent of total data were used for training and 30 per cent for testing) over the total of 4,875 observations. These data were trained with a J48[155] decision tree to establish rule delineation that hierarchically maps and associates the factors that made the market volatile. The rule delineation made using the decision tree was established using these criteria in WEKA[156], as tabulated in Table 4.13.

Table 4.13 J48 Decision Tree Classification Criteria

CRITERION	VALUE
binarySplits	True
unpruned	False
reducedErrorPruning	True
subtreeRaising	True

Table 4.14 shows the data used as input for the data training. Only 11 features from the 13 *oilFactor* features presented were established, interconnected and mapped on the decision tree, with the price data as the decision output. The data established from the decision tree training are presented in Table 4.14.

[155] The J48 decision tree is a type of WEKA classifier that uses C4.5 algorithms to split the attributes of training data into classes.
[156] WEKA: an open source classification toolkit for machine learning.

Table 4.14 The Input *coreTerm = oilFactor* Data Mapped and Established from the Decision Tree with *Price* as the Decision Output

NO.	VARIABLE	INPUT [*OILFACTOR*]
	$\boldsymbol{D^{T157}}$	**TOTAL DEMAND**
1	d_1	Demand
2	d_2	Consumption
	$\boldsymbol{I^T}$	**TOTAL INVENTORY**
3	i_1	Imports
	$\boldsymbol{S^T}$	**TOTAL SUPPLY**
4	s_1	Supply
5	s_2	OPEC decisions on productions
6	s_3	Production
7	s_4	Refinery
8	s_5	Stocks
	$\boldsymbol{P^T}$	**TOTAL POLITICS**
9	p_1	War
	$\boldsymbol{E^T}$	**ECONOMY**
10	e_1	Recession
11	e_2	US Dollar
12		Population
13		OECD decisions
14		Weather
15		Inventory

From the data training, it was found that the weather and inventory *oilFactors* were not established from the decision tree, because there was zero information obtained and no attributes interconnected between the *oilFactors*. A sensitivity analysis was also made to compare and validate the decision tree performance with other machine learning approaches, such as a support vector machine (SVM), Bayesian networks (BN) and a multi-layer perceptron (MLP). The results of this analysis are discussed in section 4.6.2.

[157] T is the accumulated value of each *oilFactor* feature.

4.6.2 Simulation Results from the Decision Tree and Other Approaches

After running all four sets of training and testing based on the data provided by the sentiment analysis, the best training result was obtained from the data that were trained with a 90:10 percent ratio and a root mean squared error (RMSE) value of 0.372, and 73% accuracy. The simulation was made in order to establish rule delineation based on the data obtained from the sentiments analysed in section 4.5. This is also to hierarchically map and associate the impact factors that caused the market to be volatile systematically.

The decision tree was also used to validate the linguistic data obtained and to compare it with the key factors found in the hierarchical conceptual (HC) model, presented in Figure 3.3 and discussed in section 3.2 of Chapter 3. It was perceived that, by providing more data for training purposes, the result obtained is more accurate. This is because more learning could be applied and achieved from the events that occurred, and which contain knowledge rules that interpret the market's behaviour, as mapped on the decision tree.

The decision tree simulation, which was compared to other machine learning approaches, had produced fairly satisfactory results based on its minimum RMSE value. Figure 4.8 shows the J48 decision tree that was produced based on the analysed sentiments gained from the extracted fragments of an article. The training that was used to produce the decision tree in Figure 4.19 was based on the 90:10 percent data of the training: testing percentage ratios. To evaluate the performance of the decision tree, an evaluation using other classification methods in machine learning approaches, such as a support vector machine (SVM), Bayesian networks (BN) and a multi-layer perceptron (MLP), was carried out with the trainings applied via WEKA. The results from these simulations are recorded in Table 4.15.

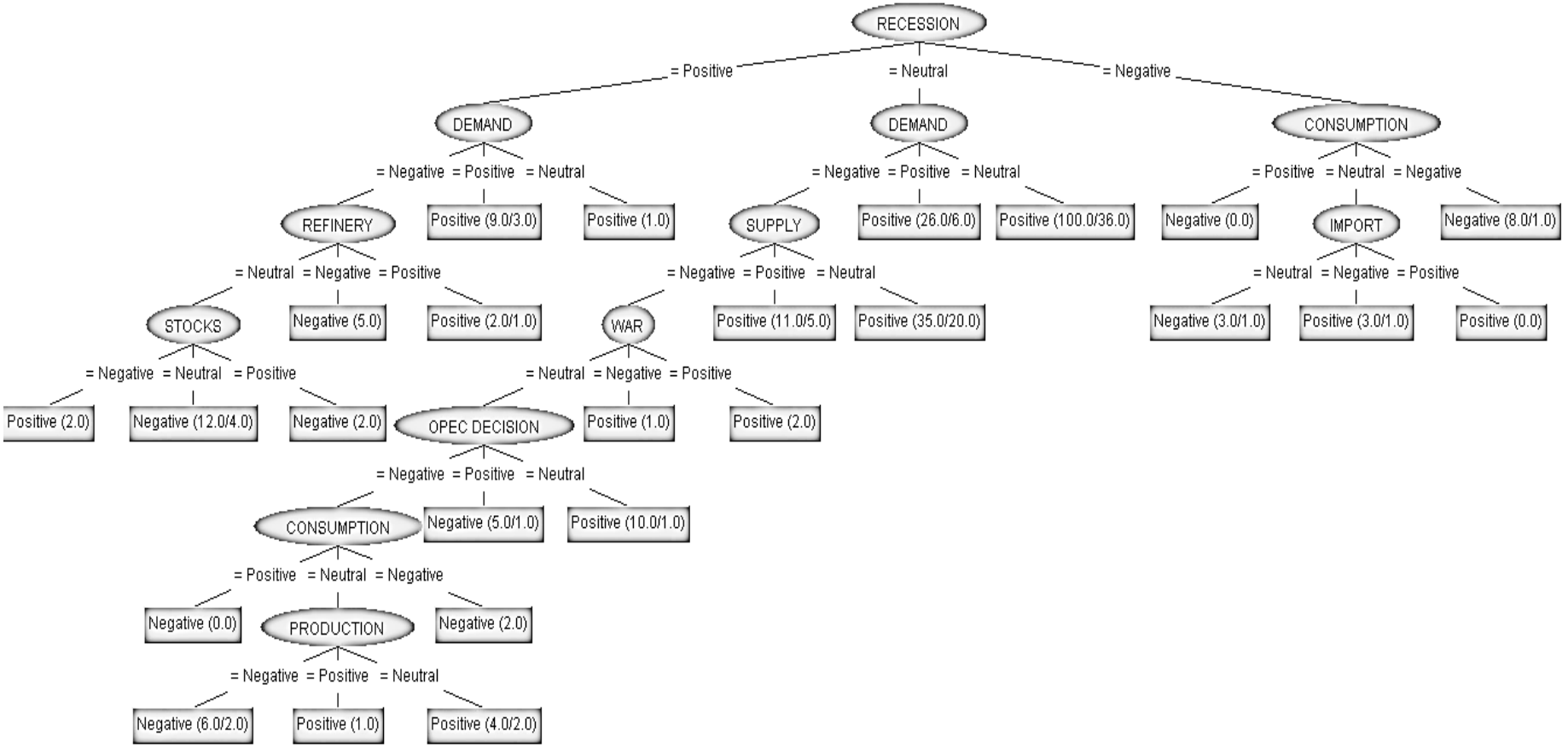


Figure 4.8 Rule Delineation Performed with the J48 Decision Tree for the Rule-Based Expert Model

Table 4.15 Results of Rule Delineation and its Comparison with other Machine Learning Approaches.

METHOD	RMSE	ACCURACCY (%)
Decision Tree		
90:10	0.372	73
80:20	0.427	64
70:30	0.450	56
Support Vector Machine	0.358	78
90:10	0.415	61
80:20	0.423	60
70:30		
Bayesian Networks		
90:10	0.360	76
80:20	0.419	61
70:30	0.439	60
Multi-Layer Perceptron	0.384	73
90:10	0.456	56
80:20	0.456	61
70:30		

RMSE: Root Mean Squared Error

Table 4.15 shows that the decision tree shares the same accuracy value as MLP, although it has a better value of RMSE. Nonetheless, SVM leads to the best result for RMSE and for the percentage of accuracy, with 0.358 and 78% accuracy, respectively, followed by BN, the decision tree and MLP. The decision tree, with its competitive result, has an advantage over the other methods as a result of its classic transparent character in classifying attributes assigned to the data training and representing them into a tree diagram. This is a very useful characteristic for mapping the rules. In the next section, the rules delineated according to this decision tree are transformed and implemented into a schematic system for the development of a fuzzy expert model.

4.7 FUZZY EXPERT SYSTEM

A fuzzy expert system is a decision-support system that emulates the reasoning process of a human expert, where expert knowledge is exploited into a systematic rule model. A fuzzy expert system consists of four main components: i) the fuzzy rule base; ii) fuzzy inference engine; iii) fuzzification; and iv) defuzzification. Ghallab, et al. [84] developed a fuzzy expert system for petroleum fields to analyse oil well data in order to motivate petroleum engineers to detect areas for reordering drilling at a new petroleum oilfield. The factors that were used as input data in the research [84] were temperature, pressure, crude oil density, gravity and gas density. The research collects accurate results on its 27 over 30 oil wells for testing. A fuzzy time-series combining the human's subjective view with objective historical values were used in the development of the expert system in [85]. Focusing on the short-term oil forecasting, research in [85] used the daily WTI price to predict the movement of prices in the crude oil market. The research obtained a good forecasting result based on the minimum RMSE value obtained from the approach.

In this chapter, fuzzy logic is used to exploit rules delineated from the decision tree in section 4.6 into a systematic rule and expert system. The rules derived from the decision tree discussed in section 4.6 were utilised as input in the fuzzy expert system through an implementation of the rule-based expert model, as discussed in this section. The development design of the fuzzy expert system incorporates the process of:

i) defining the linguistic variables obtained from section 4.5;
ii) determining the fuzzy sets—each *oilFactor* used in the fuzzy expert model was assigned to a fuzzy set, derived from the linguistic membership value obtained in section 4.6;
iii) constructing the fuzzy rules, as represented in a matrix form of 10x1 (10 inputs and 1 output variable);
iv) encrypting (ii) and (iii) into the expert system by performing the fuzzy inference; and
v) evaluating the system.

These processes are summarised in Figure 4.9 and are discussed thoroughly in the next sections.

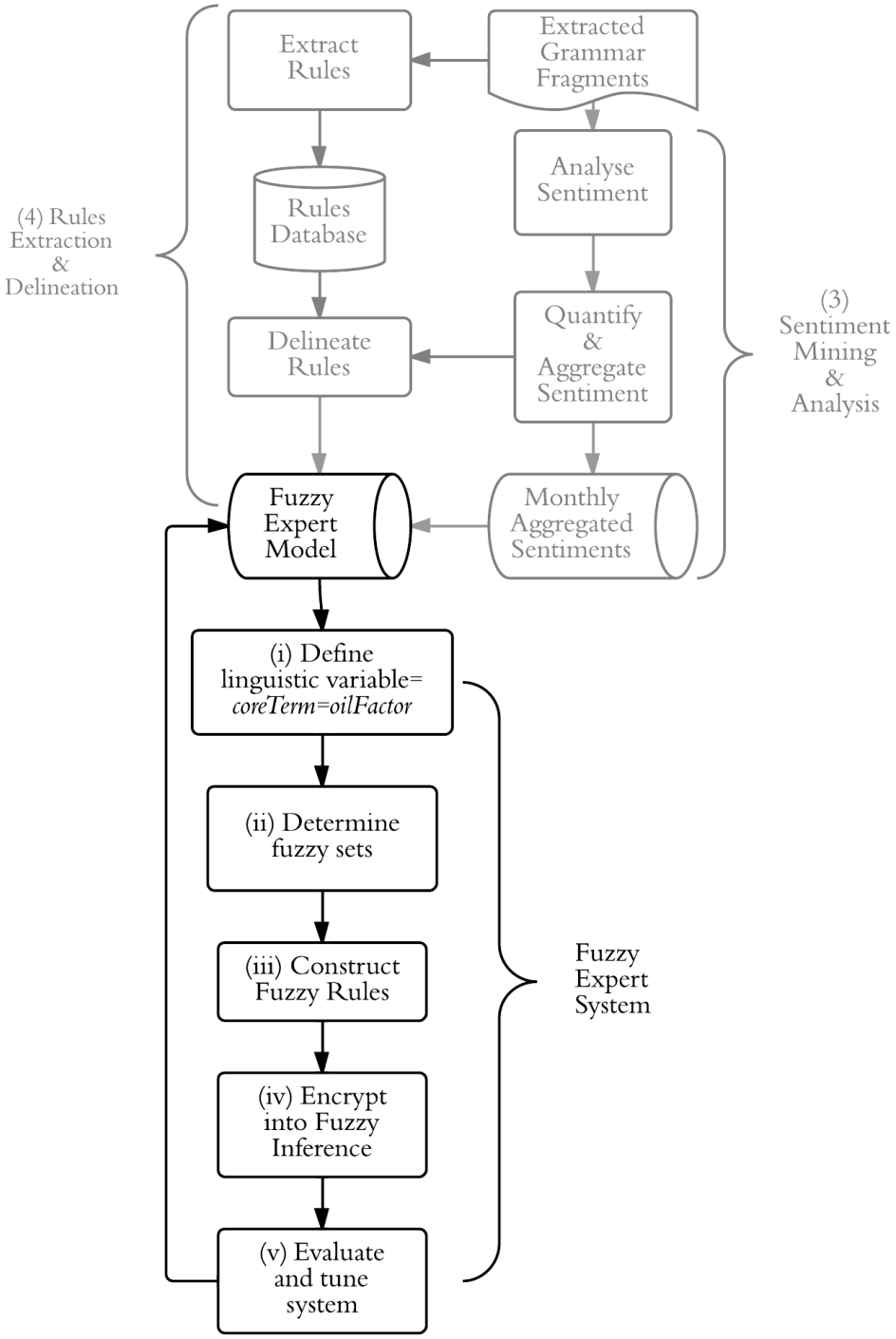


Figure 4.9 The Development Framework of the Fuzzy Expert System through a Fuzzy Expert Model.

4.7.1 Linguistic Variables and Fuzzy Sets

Preliminarily, 11 linguistic variables[158] derived from the *oilFactor* features discovered in subsection 4.6.1 and presented in Table 4.14 were assigned to and utilised in the system. These variables were validated against the crude oil market's key contributors, obtained from the hierarchical conceptual (HQ) model discussed in section 3.2 and illustrated by Figure 3.3 in Chapter 3. These profound, key factors were used in the research as a guide and as a platform for understanding the volatility that occurred in the crude oil market. The simulations made with the decision tree in section 4.6.2 validate the key factors obtained in the HQ model, with a good minimum root mean squared error (RMSE) value of 0.372. Although the percentage of accuracy gained from the experiment was only fairly satisfactory, 73%, the result is statistically significant, evidenced by its *p* value[159] = 0.01967 (significance level[160]≤ 0.05) with the potential to improve, if a larger set of data was introduced into the decision tree classifier. Initially, Fisher's exact test[161] [86] [87] [88] was established to test the significance level of the result against the null hypothesis[162]. The analysis was made on the tested data's confusion matrix[163] obtained from the decision tree simulation established in WEKA. The confusion matrix obtained from the simulation test is presented in Table 4.16 as below:

Table 4.16 J48 Decision Tree Confusion Matrix

	Positive	Negative	Neutral
Positive	20	2	0
Negative	7	7	0
Neutral	1	0	0

In Table 4.16, each column in the matrix represents the predicted value for each class with each row in the matrix representing the actual value of each class. For example, from the confusion matrix results:

[158] Linguistic variables are used to describe a term or a concept with vague or fuzzy values.
[159] *P* value: a significance level of test against/to reject null hypothesis.
[160] Significance level: a fixed probability of wrongly rejecting the null hypothesis, if it is in fact true.
[161] Fisher's exact test: a statistical significance test to analyse confusion matrix.
[162] Null hypothesis: represents a theory that being put forward, either because it is believed to be true or because it is to be used as a basis for argument, but has not been proved [114].
[163] Confusion matrix: a specific table that interpret the performance of an algorithm.

i) 20 of the total prices were correctly predicted as *positive* with 7 and 1 predicted as *negative* and *neutral;*

ii) 7 of the total prices were correctly predicted as *negative* with 2 predicted as *positive* and 0 as *neutral;* and,

iii) 0 prices were correctly predicted as *positive*, *negative* and *neutral*.

The linguistic variables obtained from the aggregated sentiments discussed in section 4.6.1 were used as the membership function in the fuzzy expert model. By referring to Table 4.11, the monthly aggregated sentiments were also derived fuzzy sets[165] that can be used as input in the inference system. These fuzzy set derivations were obtained from equations (4.4), (4.5) and (4.6), and examples of the results of this calculation are shown in Table 4.11 as purple cells. The fuzzy sets were obtained by means of the degree of membership function μ, as *positive*, *neutral* and *negative* membership, with F^{pos}, F^{neut} and F^{neg} denoting the fragments with a positive sentiment, a neutral sentiment and a negative sentiment, respectively, for *oilFactor A*, while the variable S^A denotes the monthly aggregated sentiment for each *oilFactor A*.

$$\mu_A^{positive} = \sum F^{pos} \text{ x } \left(\frac{1}{S^A}\right) \tag{4.4}$$

$$\mu_A^{neutral} = \sum F^{neut} \text{ x } \left(\frac{1}{S^A}\right) \tag{4.5}$$

$$\mu_A^{negative} = \sum F^{neg} \text{ x } \left(\frac{1}{S^A}\right) \tag{4.6}$$

[165] Fuzzy sets are sets with fuzzy boundaries, such as *positive*, *neutral* and *negative*, where each boundary is aligned with a degree of membership.

Trapezoidal and triangular fuzzy sets are often used in a fuzzy inference[166] system because of the adequate representation of expert knowledge and this often simplifies the process of computation [26]. A fuzzy inference process includes:

i) the fuzzification[167] of input variables;
ii) rule evaluation;
iii) the aggregation of output rules; and
iv) the defuzzification[168] stage.

Figure 4.10 shows an example of normalised fuzzy sets derived from the membership functions of an input=*oilFactor*=Demand, where the linguistic variables were obtained from the sentiment analysis discussed in section 4.5. Meanwhile, the fuzzy sets shown in Figure 4.11 belong to the output=*price* category.

[166] Fuzzy inference: a process to map a given input to an output.
[167] Fuzzification: a step to set numerical values (degree of membership) for linguistic variables.
[168] Defuzzification: a step to produce a crisp final output.

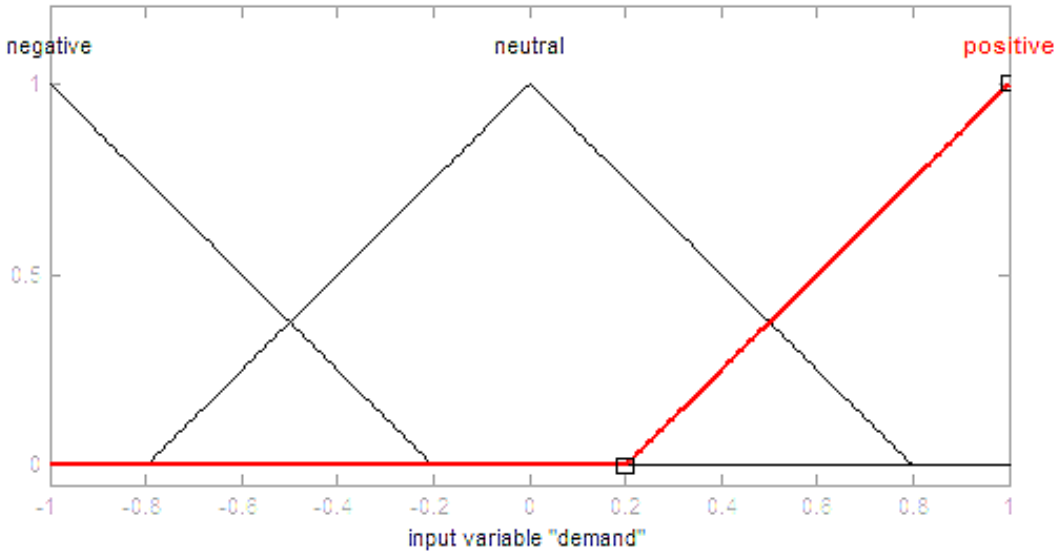


Figure 4.10 Example of Normalised Fuzzy Sets for Input *oilFactor*=Demand.

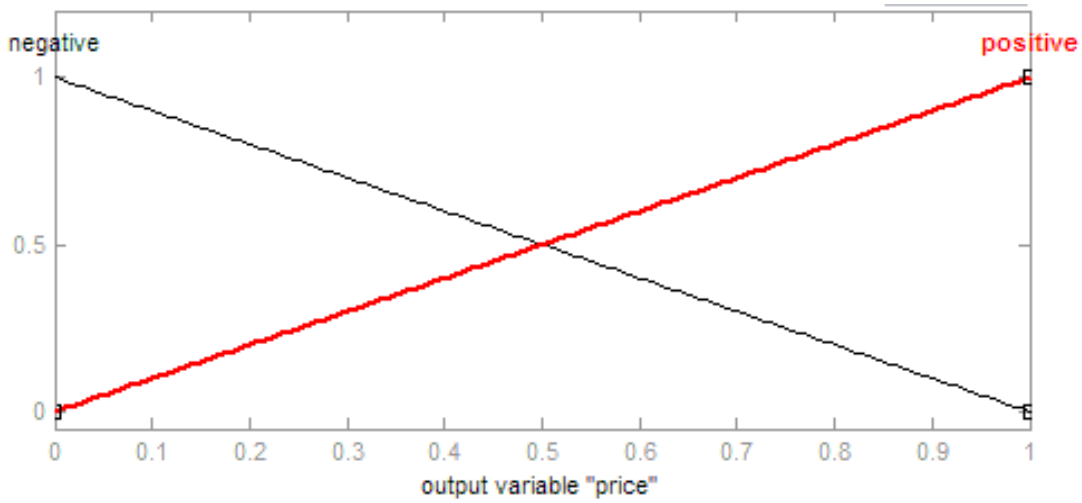


Figure 4.11 Example of Normalised Fuzzy Sets for Output=*Price*.

4.7.2 Constructing Fuzzy Rules

The input and output used for this process were obtained from the rule delineation process made using the decision tree in section 4.6.1, which were then applied to a fuzzy inference system in order to produce a schematic form of application. In this section, the language variables and prices used as input and output in the fuzzy expert model were represented as fuzzy rules in a matrix form of a 10-by-1 system[169]. Fuzzy rules[170] were used to capture the domain knowledge presented by the language variables and the fuzzy sets discussed in section 4.7.1.

Consequently, 25 fuzzy rules were derived for the price representing the complex relationships between all variables used in the expert system. Figure 4.12 shows the rules that were obtained from the utilisation of the membership functions discussed in sections 4.5 and 4.6 and represented in verbose format.

[169] 10-by-1 system: Ten linguistic variables are depicted as input and price as one output.
[170] Fuzzy rule: A fuzzy rule is a condition statement in the form: IF *x* is *A*, THEN *y* is *B.*

Rule Base 1: Price

1. If (recession is positive) and (demand is negative) and (refinery is neutral) and (stock is negative) then (price is positive)
2. If (recession is positive) and (demand is negative) and (refinery is neutral) and (stock is neutral) then (price is negative)
3. If (recession is positive) and (demand is negative) and (refinery is neutral) and (stock is positive) then (price is negative)
4. If (recession is positive) and (demand is negative) and (refinery is negative) then (price is negative)
5. If (recession is positive) and (demand is negative) and (refinery is positive) then (price is positive)
6. If (recession is positive) and (demand is positive) then (price is positive)
7. If (recession is positive) and (demand is neutral) then (price is positive)
8. If (recession is neutral) and (demand is negative) and (supply is negative) and (war is neutral) and (opec_decision is negative) and (consumption is positive) then (price is negative)
9. If (recession is neutral) and (demand is negative) and (supply is negative) and (war is neutral) and (opec_decision is negative) and (consumption is neutral) and (production is negative) then (price is negative)
10. If (recession is neutral) and (demand is negative) and (supply is negative) and (war is neutral) and (opec_decision is negative) and (consumption is neutral) and (production is positive) then (price is positive)
11. If (recession is neutral) and (demand is negative) and (supply is negative) and (war is neutral) and (opec_decision is negative) and (consumption is neutral) and (production is neutral) then (price is positive)
12. If (recession is neutral) and (demand is negative) and (supply is negative) and (war is neutral) and (opec_decision is negative) and (consumption is negative) then (price is negative)
13. If (recession is neutral) and (demand is negative) and (supply is negative) and (war is neutral) and (opec_decision is positive) then (price is negative)
14. If (recession is neutral) and (demand is negative) and (supply is negative) and (war is neutral) and (opec_decision is neutral) then (price is positive)
15. If (recession is neutral) and (demand is negative) and (supply is negative) and (war is negative) then (price is positive)
16. If (recession is neutral) and (demand is negative) and (supply is negative) and (war is positive) then (price is positive)
17. If (recession is neutral) and (demand is negative) and (supply is positive) then (price is positive)
18. If (recession is neutral) and (demand is negative) and (supply is neutral) then (price is positive)
19. If (recession is neutral) and (demand is positive) then (price is positive)
20. If (recession is neutral) and (demand is neutral) then (price is positive)
21. If (recession is negative) and (consumption is positive) then (price is negative)
22. If (recession is negative) and (consumption is neutral) and (import is neutral) then (price is negative)
23. If (recession is negative) and (consumption is neutral) and (import is negative) then (price is positive)
24. If (recession is negative) and (consumption is neutral) and (import is positive) then (price is positive)
25. If (recession is negative) and (consumption is negative) then (price is negative)

Figure 4.12 A Fuzzy Rule Base for Price

4.7.3 Building and Evaluating a Fuzzy Expert System

According to [89], the core components of expert systems include the knowledge base to store factual and heuristic knowledge. The knowledge gained from the decision tree discussed in section 4.6.1 was utilised in this fuzzy expert model and presented in Figure 4.16, and stored as IF-THEN rules. Another factor is a reasoning engine, which utilised the stored knowledge mentioned above in an inference system as a reasoning engine to form a line of reasoning rules. In addition, a knowledge acquisition subsystem is an important base for executing the knowledge stored in the IF-THEN rules in order to build the expert system. Finally, an explanation subsystem that functions as a platform for evaluating the system is required. The core components of the first two points were discussed in subsections 4.8.1 and 4.8.2 as a preparation phase for developing the expert system. The remaining core components will be discussed in this section as the development phase of building the expert system.

To accomplish the development phase, the fuzzy sets and fuzzy rules were encrypted to perform a fuzzy inference in order to build the expert system. In this section, the Mamdani method was used as fuzzy inference instead of the Sugeno method. The crisp output generated by the Mamdani method was based on the defuzzification technique, while the Sugeno method uses a weighted average computation. The expressive power and interpretability of the Mamdani output is lost in the Sugeno fuzzy inference system, since the consequents of the rules are not fuzzy [90]. Due to the interpretable and intuitive nature of the rule base, the Mamdani fuzzy inference system was chosen because it is widely used, particularly for decision-support applications, as suggested by [91].

The fuzzy expert system in this research was built by applying the fuzzy logic development tool MATLAB Fuzzy Logic Toolbox® from MathWorks [92]. The fuzzy logic development tool was used as the reasoning engine to provide complete environments for building and tuning the fuzzy system. Since building a fuzzy expert system is an iterative process that involves tuning the existing fuzzy sets and fuzzy rules to meet specific requirements, applying a fuzzy development tool would help to speed up the tuning process. In order to analyse the rules' performance, surface

graphs were generated to visualise the rules and are represented by three-dimensional plots in Figure 4.13 and the rules of the fuzzy expert system based on the Mamdani-type fuzzy inference system are presented graphically in Figure 4.14.

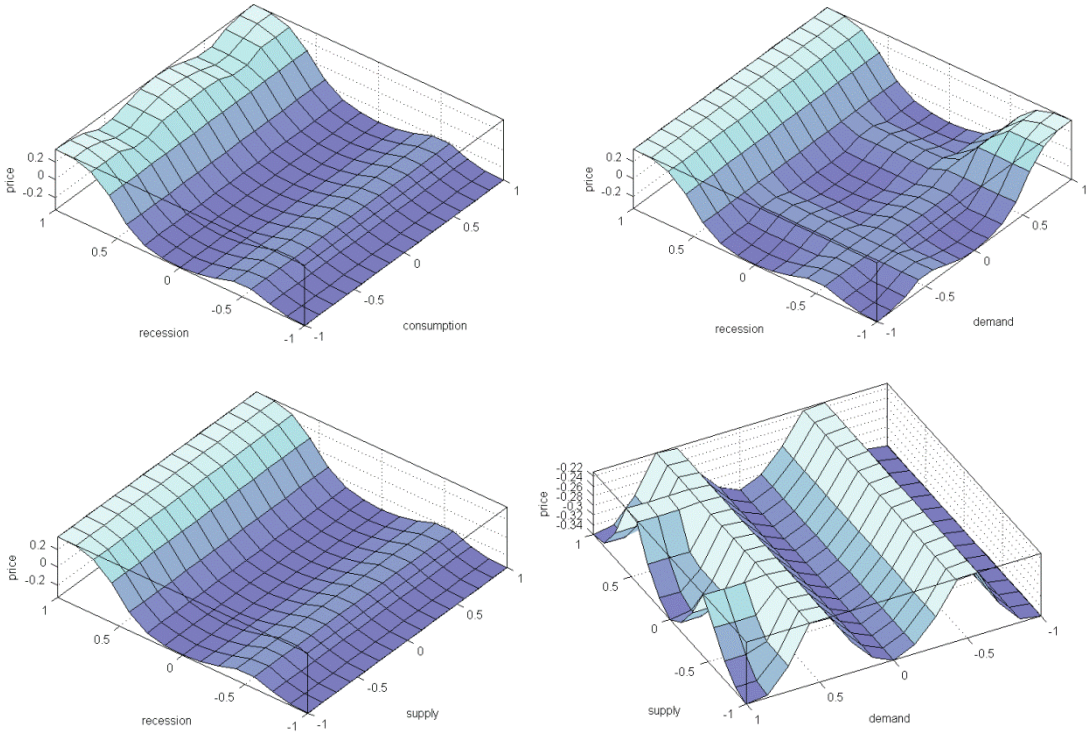


Figure 4.13 Example of Inference System Output Surface View.

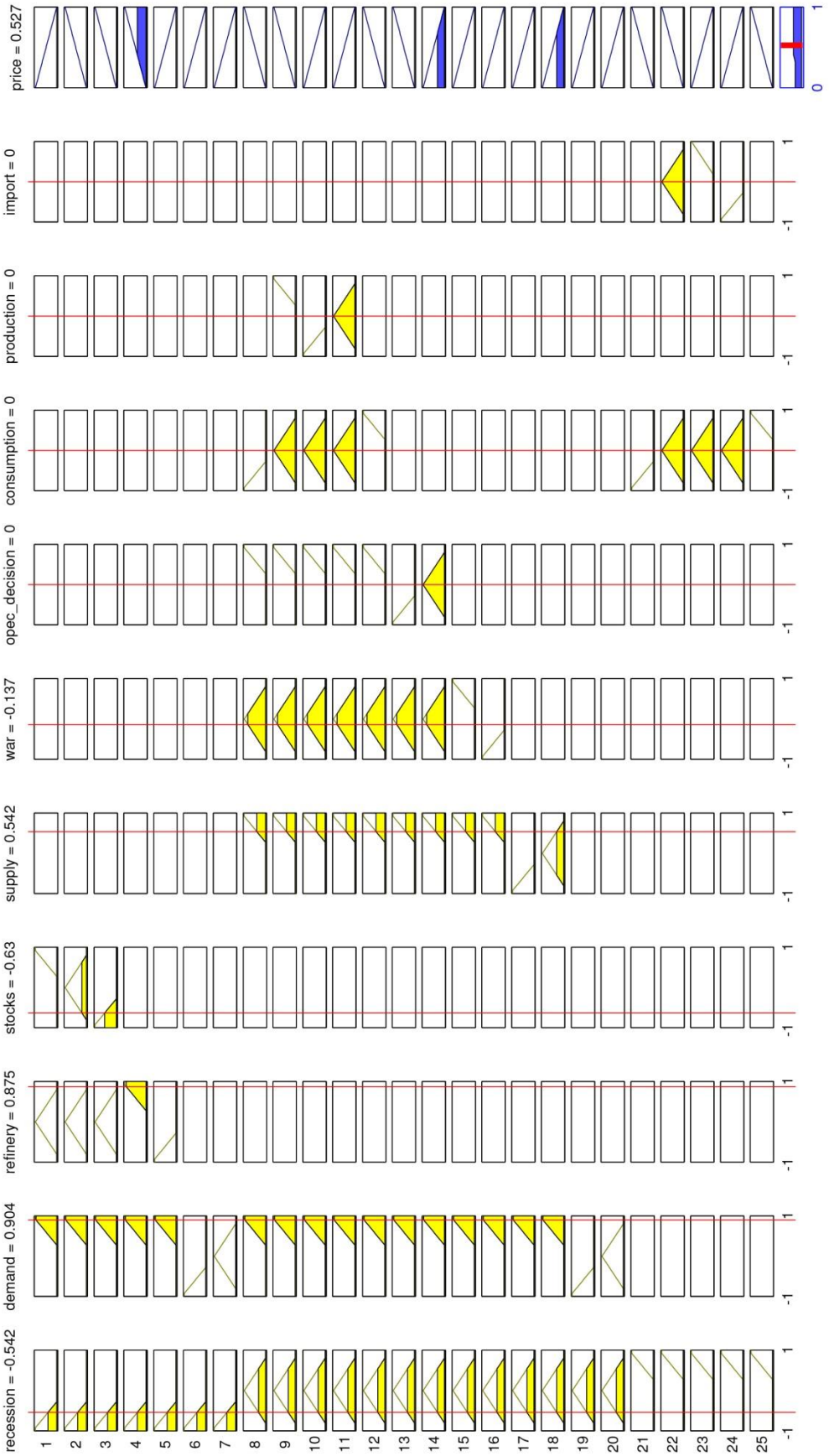


Figure 4.14 Example of Rule Viewer for Price.

Table 4.17 The Rule Evaluation with the Fuzzy Expert System

NO.	LINGUISTIC VARIABLES	FUZZY SETS
1	Recession	*Negative*
2	Demand	*Positive*
3	Refinery	*Positive*
4	Stocks	*Negative*
5	Supply	*Positive*
6	War	*Negative*
7	OPEC Decisions	*Neutral*
8	Consumption	*Neutral*
9	Production	*Neutral*
10	Import	*Neutral*
	Price	**0.527 (Positive)**

According to Figure 4.14, Table 4.17 exhibits the evaluation made for the set of fuzzy rules implemented in the fuzzy expert system. The 0.527 crisp degree shown in Table 4.17 for the crude oil price was manifested from the multiplication degrees of "Recession", "Demand", "Refinery", "Stocks", "Supply", "War", "OPEC Decisions", "Consumption", "Production", and "Import" as the *oilFactor* degrees to the *price*; with its 0.527 crisp degree indicates the 52.7% of positive price increase. The parameters used in the fuzzy inference system for the fuzzy rule evaluations were based on the example values set in Appendix B, which were induced in MATLAB.

4.8 CONCLUSION

In this chapter, the development of a rule-based expert model that involved the utilisation of Google News service content for grammar building through fragment extraction and grammar derivation was discussed, as was sentiment analysis. The process of constructing a grammar definition based on language structure with Context-Free Grammar (CFG) was discovered. The definitions that were set up in a terminal grammar to derive grammars in extracted text fragments were utilised. In the grammar derivation process, a grammar combination of a *coreTerm* and a *contentCategory* in a fragment from an article was successfully annotated and extracted. The sentiments of these extracted grammar fragments, which carry expert domain knowledge, were analysed, quantified into -1, 0, and 1, and aggregated. This chapter contributed to producing new input data, which were derived from the aggregated sentiments and utilised in the rule-based expert model. The new inputs encoded in the extracted grammar fragments in the grammar derivation process constitute knowledge that is to be employed in the linguistic prediction model in the next chapter. The quantified and aggregated sentiments were also utilised as input in the decision tree in order to map the rules. The decision tree successfully mapped the knowledge behaviour and reasoned correctly, with 73% accuracy and a 0.372 root mean squared error (RMSE). The decision tree's percentage of accuracy is statistically significant, evidenced by its p value= 0.01967 (significance level ≤ 0.05) with the potential to improve if a larger set of data was introduced into the decision tree classifier. The tree produced with this procedure was then used to derive the fuzzy sets and fuzzy rules in the fuzzy inference system as input for forming the rules systematically and for establishing the expert system.

This chapter produced core contributions to this research by first successfully defining the grammar definitions and building the grammars that are specifically tailored to the crude oil market. Importantly, these definitions helped to derive appropriate grammar fragments from news articles, which enabled the discovery and the analysis of the sentiments from each extracted fragment. These sentiments were not only useful for the analysis but also for constructing the decision tree and the linguistic variables for the fuzzy expert system. Later in this chapter, 25 rules are discovered and used as a rule base to evaluate the price factor. The rules are

employed in a fuzzy inference system to build an expert system that can be modified and evaluated from time to time.

Chapter 5 Linguistic Prediction Model with Sentiments Mined from Google News Articles and Its Hybridisation with a Quantitative Prediction Model

Overview

Future expectations are usually derived from historic information retrieval and heuristic evaluation through a set of time-series data. Mining from the right and significant source of information with appropriate use of a keyword is consequently an important challenge. This retrieval process is implemented in order to anticipate future occasions and to aggregate it for imminent decisions. The good results obtained from the Artificial Neural Network-Quantitative (ANN-Q) model in Chapter 3 of this thesis resulted from the systematic approach introduced in the Hierarchical Conceptual (HC) model of section 3.2. An extension of this quantitative prediction model, with the contribution of linguistic features, is introduced in this chapter. A fuzzy grammar fragment extraction module using a rule-based expert model in Chapter 4 effectively derived a set of sentiments from mined Google News articles, which were exploited as rules for the fuzzy expert model in section 4.7. These rules were used as input for this linguistic prediction model. This chapter discusses the exploitation of sentiments as linguistic input for predicting the crude oil price, the development of ANN modelling to predict using the linguistic information, the hybridisation of both linguistic and quantitative data for predictions and, finally, the empirical results gained from these linguistic and quantitative models.

5.1 INTRODUCTION

Exploiting text as a source of data in order to understand financial behaviour is currently popular. The application of this text analysis for financial literature is assisted by the emergence of the Internet, where vast numbers of financial texts are easily obtained from online news services. The need to exploit this information is indispensable. News articles mined from the Google News service were chosen for this research in order to understand the domain behaviour within the hypothesis, as they contain significant events that rule (ie. affect) crude oil market behaviour. Cecchini, et al. in [93] demonstrated that text and quantitative information give the best results when combined. Even when benchmarked only with quantitative information, textual information has proven to be competitive in terms of prediction. This is proven by the authors' [93] examination of the role played by financial texts in predicting corporate fraud and bankruptcy, wherein this examination resulted in 82.0% and 83.9% accuracy, respectively.

The authors of [94] [95] [96] [97] also believed that information extracted from news articles could have a visible impact on price. Nevertheless, Schumaker, et al. [98] believed that associating news articles with quantifiable price movements is still a difficult task, specifically when predicting the behaviour of a stock market. Tetlock [99] also found that qualitative information from the text complements quantitative information. Tetlock, et al. [100] observed that the text not only indicates the events of a domain study, but also adds value to the indicated events. Meanwhile, Das and Chen [101] extended the study [102] by using stock price message boards to measure bullishness in the stock market. Although the findings from both of the studies show weak and noisy predictive ability, they discovered that even message boards carry valuable information. Davis, et al. in [103] measured the effect of the tone of a body of text on financial outcomes and discovered that an optimistic tone contributes to a higher future return of assets, which affects share prices. This study [103] also confirmed the important contribution of textual information to a quantitative prediction model.

The quintessence of the studies mentioned [93]- [103] is that textual information in terms of news articles from the Google News service is perceived to share the same sentiment. Following the pattern of the various domains mentioned, interesting results will be gained from this hybridisation of linguistic and quantitative prediction research.

5.2 THE LINGUISTIC INPUT: SENTIMENT MINING FROM THE GOOGLE NEWS ARTICLES

The process of sentiment mining and analysis has been discussed in detail in the sentiment analysis section in Chapter 4. Through this process, we discovered expert knowledge that was later applied to an expert model. In Chapter 4, we discovered expedient sentiments mined from the news fragments, which count as the essential elements in this linguistic model.

5.2.1 Sentiment Mining

Discovering the sentiment of fragments extracted from an article in a periodical provides an overview of the crude oil market, by analysing its patterns of behaviour and understanding its volatile contributions. Through the analysis of periodicals, one can understand the past and anticipate future events, as well as utilise decisions for the next period selected. This explains the importance of the sentiment mining process in influencing future anticipation. Figure 5.1 shows the sentiment mining process according to a linguistic prediction model framework. It shows the mined and analysed sentiments from fragmented articles, which are quantified and aggregated monthly, are then used as data input in this linguistic prediction model. These monthly sentiments are simply quantified to allow for this linguistic data to be implemented by utilising an artificial neural network (ANN), which is discussed in section 5.3. The input is employed in the linguistic prediction model in order to recognise the crude oil market's pattern of behaviour according to the positive or negative tone discovered in a fragment, which is extracted from a statement in an article. The employment of the sentiment input as the main data for this linguistic prediction model will propagate directional connections between inputs in the model by calculating the 'risks' (weights) between the linguistic inputs and, later, will produce a directional outcome in terms of positive or negative directions from the predicting module.

Previously, subsection 4.3.4 of Chapter 4 discussed the crucial process of constructing a customised 'dictionary'. The 'dictionary' consists of a set of grammars, called terminal grammar, which is set up as the first process to be prepared before extracting the fragments from the statements of an article. The

terminal grammar, through the application of fuzzy grammar fragment extraction mentioned in section 4.4 of Chapter 4, constructs a word-tagging process in the fragment extraction module. In the process, each word in an extracted fragment is tagged with a customised 'definition', or category, that was defined earlier in the terminal grammar. Fragments were extracted based on a grammatical combination of words in the article that contain a defined 'noun', such as *oilFactor,* with its positive or negative 'verb', *positiveTerm* or *negativeTerm*. From this sentiment-mining process, approximately 5,250 grammar fragments were extracted for the period from January 2008 until December 2009. These grammar fragments were allocated a sentiment according to equation (4.1) and equation (4.2), as stated in section 4.6 in Chapter 4. These sentiments were then aggregated into months in order to enable a sense of the monthly sentiment, because the objective of the linguistic prediction model is to predict the direction of the next month's price. We later extrapolated these monthly sentiments into a linguistic prediction model in order to determine its capacity to predict the monthly crude oil price via the linguistic entities.

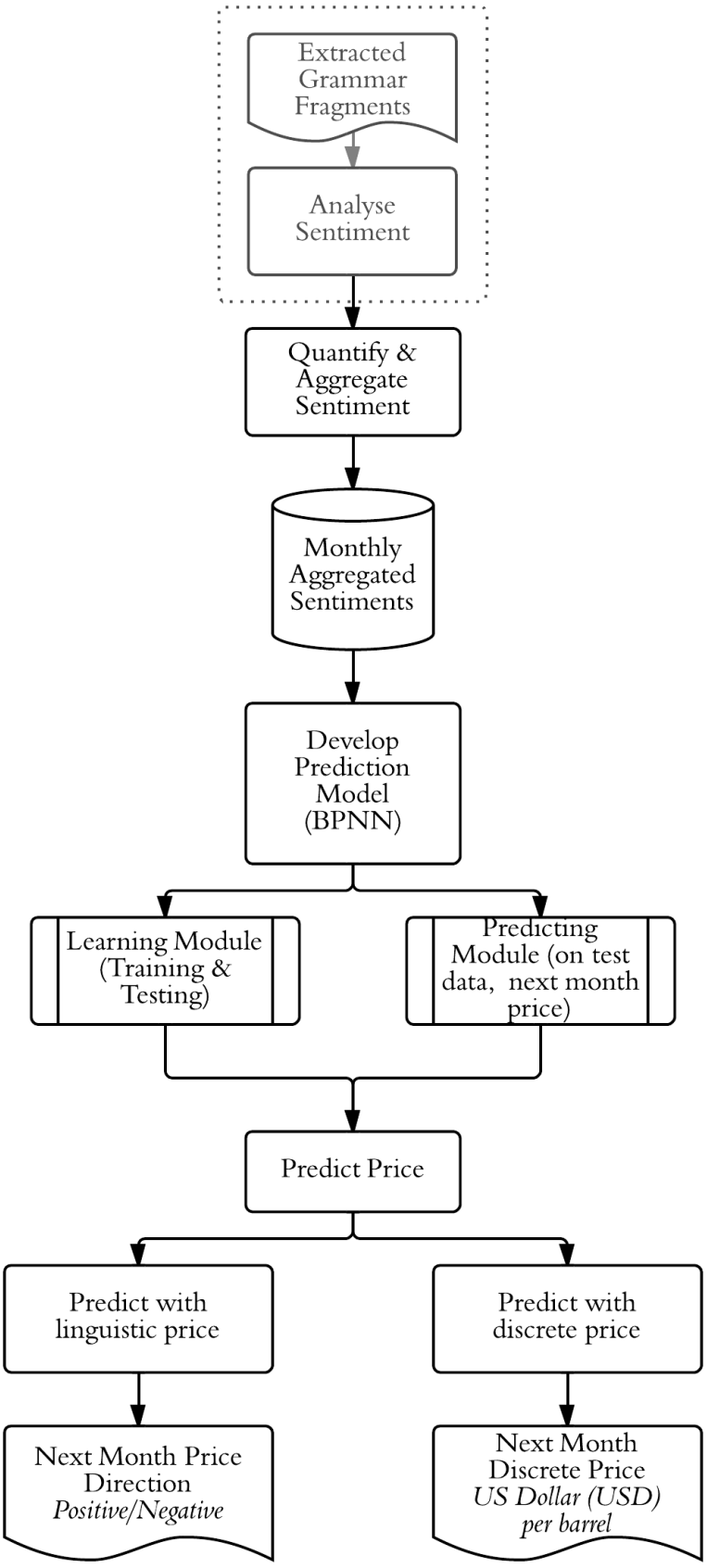


Figure 5.1 Sentiment Mining and Linguistic Prediction Model Framework.

5.3 THE ARTIFICIAL NEURAL NETWORKS (ANN) LINGUISTIC PREDICTION MODEL

ANN is well known for its capacity for capturing the nonlinear dynamics of the crude oil price market. Xie, et al. [36] observed that it outperformed most other methods in their study. Moshiri and Foroutan [104] demonstrated that ANN effectively improved the performance of modelling the unknown nonlinear economic structure in the authors' studies of crude oil futures prices. This is also supported in the work by Shambora and Rossiter [105]. The authors attested that ANN delivers superior performance for the measurement of its cumulative returns over market cycle returns and Sharpe ratios. Despite its 'black box' and unstable character highlighted by[24] [41] [106], ANN's performance in modelling the nonlinearity of data is still impressive and is relevant for further exploration through the implementation of a linguistic prediction model and a prediction using the hybridisation of both linguistic and numeric data.

In this section, linguistic information is utilised as input in the next stages of crude oil price prediction. To evaluate its capabilities in predicting the crude oil price, the linguistic information gained from sentiment mining and analysis was employed in two models:

i) The linguistic prediction model validates the use of linguistic input obtained from Chapter 4 as a credible data input that influenced the behaviour of the crude oil market. The linguistic features that constitute the sentiment of an extracted fragment are quantified, aggregated and are used as the monthly linguistic input for predicting the market according to three different types of prices —directional (prices are quantified as 1 = positive and 0 = negative), normalised (prices are quantified in between -1 to 1) and original price (normal prices to two decimal places) in section 5.4 of this chapter; and,

ii) The linguistic-quantitative (LQ) prediction model complements the quantitative prediction model with linguistic features for a study to improve the quantitative prediction model's performance and to examine the credibility of the linguistic features in adding value to the quantitative prediction model.

5.3.1 Back-Propagation Trained Neural Network (BPNN) Learning Algorithm

The back-propagation (BP) learning algorithm aims to minimise error and consists of two processes, as presented in Figure 5.2, namely input training and back-propagation errors. The training input and the training output in the first process are normally stored in two matrices. This input training process is a feed-forward phase that clamps the input vector *X* to the ANN input and then propagates it through the network to calculate *Y*, where

$X = (x_1, x_2, \dots, x_n)$ represents the training input vector of *oilFactor* and $Y^t = (y_1^t, y_2^t, \dots, y_m^t)$ denotes the training to set (*t*) the output vector for price,

where *n* is the number of neurons in the input layer and *m* is the neurons for the output layer.

Initially, the input pattern in *X* is propagated through the network from layer to layer until the output activation of the pattern is generated at the output layer. The inputs in *X* are first clamped to the network's input layer in order for the trained layer to recognise the patterns contained in the data. By referring to Figure 5.2, *X* is propagated from left to right in (i) to signal its input forward through the network.

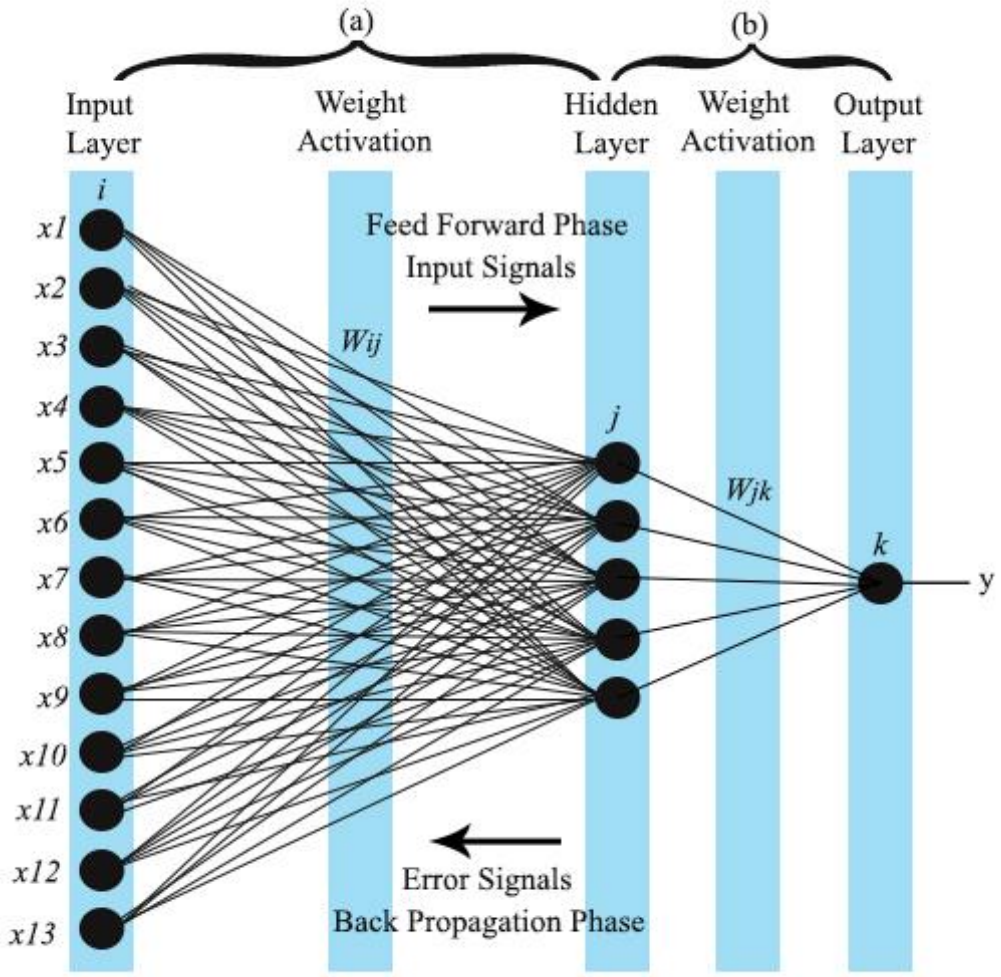


Figure 5.2 A Three-Layer Artificial Neural Network (ANN) Trained with Back-Propagation (BP) Topology in a 13-5-1 Structure.

The propagation begins when *X* propagates its *oilFactor* from the input on the input layer to the hidden layer, to finally activate the hidden layer output, y_j. This hidden output, y_j, is forward propagated as an input to the hidden layer in order to generate the final output y_k, the predicted price output of the output layer. The forward propagated activation is computed by the inter-layer sequence (left to right) presented in Figure 5.2 at the feed-forward phase in a), the input layer weight-activation hidden layer and b), the hidden layer-weight activation-output layer order. At (b), the hidden layer actually activates the output generated from (a) and forward propagates it to (b) as the input. The calculation of the output activation for each layer is calculated as:

$$y_j = f_j\left[\sum_{i=1}^{n} x_i w_{ij}\right] - \theta_j \tag{5.2}$$

$$y_k = f\left[\sum_{j=1}^{m} y_j W_{jk}\right] - \theta_0 \tag{5.3}$$

with *w*, θ, *i*, *j* and *k* denoting the weight, threshold and neurons for each input and output layer, respectively. Meanwhile, *ij* and *jk* are the connections between the hidden-input layer and the hidden-output layer. The error is a function of outputs y_j and y^t; outputs y_j and y_k use the *f* functions in equation 5.4 to activate the BP in order to obtain a satisfactory error criterion from the network. The functions used as the activation are the log sigmoid, the zero-based log sigmoid and the hyperbolic tangent, presented respectively in equation 5.4. The implementation of different functions in the network is also intended to establish a performance comparison for each function when associating the data.

$$Y^{log\ sigmoid} = \frac{1}{1+e^{-x}}$$

$$Y^{zero-based\ \log\ sigmoid} = \frac{e^{a+bx}}{1+e^{a+bx}}$$

$$y_j \text{ and } y_k = Y^{tan\,h} = \frac{2a}{1+e^{-bX}} - a \tag{5.4}$$

While the two former functions in equation 5.4 are the common activation functions in ANN, the latter is employed in this research as an accelerator in order to examine the acceleration contribution, so as to improve the output error minimisation. Values *a* and *b* in equation 5.4 represent the constants of the network, which are suitably given values of *a* = 1.716 and *b* = 0.667, as suggested by [106].

Later, e^{sum} is obtained and recorded through the network by summing the price error *e*, which is a function of the input (*x*); accordingly

$$e^{sum} = \sum_{l=1}^{m} f(e_l) \tag{5.5}$$

where *e* is obtained from the difference in computation between trained and predicted outputs from the hidden and output layers, respectively, as

$$e = y_k^t - y_k \tag{5.6}$$

Nevertheless, $E = (e_1, e_2, \dots, e_l)$, which represents the price errors from the output layer and back propagates this to the input layer via the back propagation in phase (ii). Sequentially, a function of $f(E)$ is propagated back to the input layer through output-hidden-input in order to retune the weight and threshold values in the network. This error back propagation is done by first computing the error gradient and the weight corrections of the output layer with

$$\delta_k = y_k(1 - y_k)e \quad \text{and} \quad \Delta w_{jk} = \alpha(y_j)(\delta_k) \tag{5.7}$$

and, later, continuing with the calculation of the error gradient and weight correction of the hidden layer,

$$\delta_j = \left[y_j(1 - y_j)e\right]\left[\sum_{k=1}^{l} \delta_k(w_{jk})\right] \quad \text{and} \quad \Delta w_{ij} = \alpha(x_i)(\delta_j) \tag{5.8}$$

where, in equation 5.8, the value of δ_j is calculated by multiplying it by the sum of the error gradient and weight value, computed from the output layer. Finally, the weight is updated by

$$w_{jk+1} = w_{jk} + \Delta w_{jk} \quad \text{and} \quad w_{ij+1} = w_{ij} + \Delta w_{ij} \tag{5.9}$$

where w_{jk+1} and w_{ij+1} represent the updated weight for the output and the hidden neurons, respectively.

The algorithm continues to iterate until a smaller error is achieved. The output performances are evaluated based on the summed, squared errors produced by the network. Table 5.1 shows the 14 separate input vectors for the linguistic prediction model discussed in this chapter, which is based on the aggregated *oilFactor* values obtained from the quantified sentiments mined from the Google News service.

Table 5.1 The Linguistic Features of the Linguistic Prediction Model.

VARIABLE	INPUT (*X*) [*OILFACTOR*]	VARIABLE	OUTPUT (*Y*) [PRICE (USD)]
D^{T}[171]	**TOTAL DEMAND**	WTI^{T}	West Texas
d_1	Demand		Intermediate (WTI)
d_2	Consumption		price.
I^{T}	**TOTAL INVENTORY**	WTI_1[172]	Linguistic price.
i_1	Inventory	WTI_2	Normalised discrete
i_2	Imports		price.
		WTI_3	Original discrete price.
S^{T}	**TOTAL SUPPLY**		
s_1	Supply		
s_2	OPEC decisions on productions		
s_3	Production		
s_4	Refinery		
s_5	Stocks		
P^{T}	**TOTAL POLITICS**		
p_1	War		
W^{T}	**TOTAL WEATHER**		
w_1	Weather		
E^{T}	**ECONOMY**		
e_1	Recession		
e_2	US Dollar		

[171] T is the accumulated value of each *oilFactor* feature.
[172] WTI_1, WTI_2 and WTI_3 were predicted in separate networks.

5.4 THE LINGUISTIC PREDICTION MODEL: ITS NETWORK ARCHITECTURE AND PARAMETER

A linguistic prediction model was built to quantify the linguistic features discussed in subsection 5.2.1 and to validate the exploitation of input obtained from the extracted grammar fragments via news sentiment mining and analysis with relation to influencing the price market. It is also designed to explore the crude oil market's behavioural trend, as mentioned in section 5.1. In this section, we discuss the architectures and parameters utilised to implement the model, together with the simulation tests and empirical results for its performance evaluation. Figure 5.2 shows the framework that was built for the linguistic prediction model using BPNN through its three layers of processes nodes :

i) data pre-processing;
ii) model architecture and parameters; and
iii) the price prediction output.

The monthly aggregated inputs retrieved from the sentiment mining and analysis stage are presented as $x \in \{-1, 0, +1\}$, implying negative, neutral and positive contributions in a normalised form and are perceived to be cleaned of noise. This explains the absence of the data pre-processing process from the framework greyed in Figure 5.3.

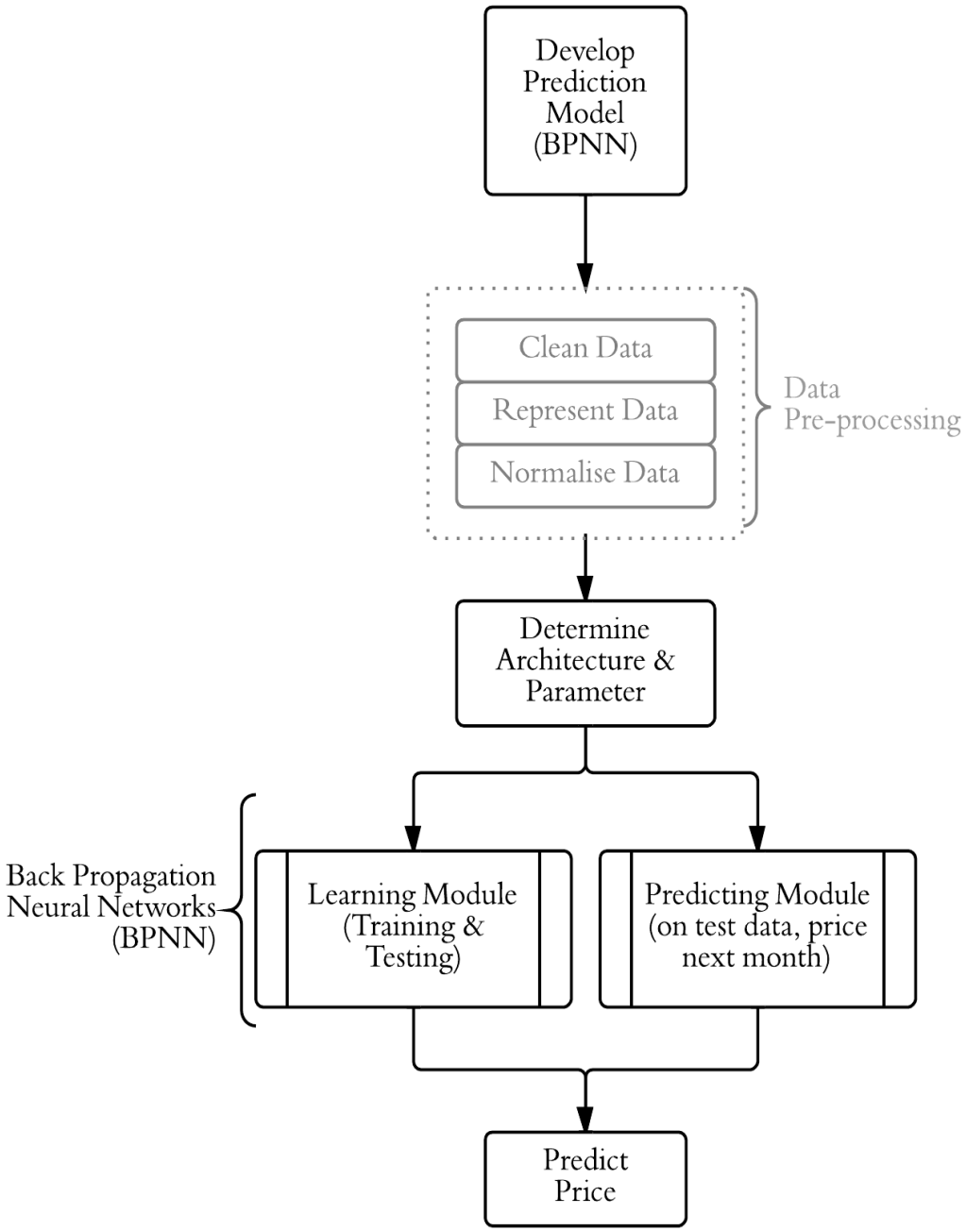


Figure 5.3 The Development of the Back-Propagation Neural Network (BPNN) for the Linguistic Prediction Model.

The architecture of this model is encoded to consist of only one input layer, one hidden layer and one output layer; a three layer ANN represented by 13-4-1 and 13-5-1 (input-hidden-output) topologies. The results and comparisons from the use of these topologies are discussed in subsection 5.4.2. Neurons for the hidden layer are randomly chosen and are trained with four and five hidden layers in order to obtain minimum errors from the linguistic prediction model. The same numbers of neurons are also applied in Chapter 3, discussed in subsection 3.3.4 and are substantiated to offer robust prediction outcomes with five hidden neurons. An example of this network topology is presented in Figure 5.3. In order to train the model, a set of parameters was chosen to optimise the process of minimising errors in obtaining accurate prediction results. The results in section 5.3.2 were obtained by utilising the parameters set in Table 5.2. These parameters provide good functions, which result in good prediction performance.

Table 5.2 Back-Propagation Neural Network (BPNN) Parameters
For the Linguistic Prediction Model.

PARAMETER	**VALUE**
Initial Weight	0.3
Learning Rate	0.3
Momentum	0.6
Learning Cycle (epochs)	5000
Network Layer	3
Neurons: Input Layer	13
Hidden Layer	4
	5
Output Layer	1
Sigmoid	Zero-based Log Sigmoid
	Log Sigmoid
	Hyperbolic Tangent

5.4.1 Sensitivity Analysis of Artificial Neural Networks (ANN) and Parameters

In order to predict the output price, the input data consists of the *oilFactor* in Table 5.1, which was used in the training and testing phases, in sequence, through the following modules:

i) the learning module—this is mathematically discussed in the algorithm mentioned in equation 5.2 as the input-hidden layer; and
ii) the predicting module (error correcting)—as the hidden-output layer.

In the learning module, the neuron was trained with three different activation functions, as follows:

a) a zero-based log sigmoid;
b) a log sigmoid; and
c) a hyperbolic tangent mentioned in equation 5.4, with four and five hidden neurons in its hidden layer.

Both (a) and (b) are often chosen to provide associations between the input data that are known to perform best with nonlinear data. Meanwhile, (c) is often employed to perform acceleration in the training process. Employing this function in the model in conjunction with the former two will, to some degree, provide adequate challenges for transmitting the prediction outcomes. The results of the activation function sensitivity analysis are shown in subsection 5.4.2.

Meanwhile, four and five neurons were chosen and were set for the hidden layer to forward- and back-propagate the weight signals from each layer. The computation accuracy of a network is often based on the selection number of the neurons in the hidden layer. Therefore, a careful consideration for this assignment is not to downgrade the network ability to perform good propagation. Larger numbers of neurons are often a burden to the network, with their heavy computational overheads and leaning towards node, learning similar activation output mapping functions. Therefore, in this research, a small number of neurons are utilised in all of the prediction models in order to minimise errors and to gain improved prediction results.

5.4.2 Sensitivity Analysis: Simulation Test and Results

Monthly aggregated sentiments mined from the Google News article service were used in this simulation test, with data that spanned from January 2008 to December 2009. This monthly data was derived from the aggregated sentiments in Chapter 4 and is perceived as having sufficient knowledge and enough interesting information for the network to learn to propagate and to then predict relevant information. This interval consists of 336 instances, which were divided into a training: testing percentage data ratio (a cross-validation of data) in the percentage order of 90:10 percent, 80:20 percent, 70:30 percent, and 60:40 percent in order to analyse the sensitivity of the data used in this linguistic prediction model. For example, 90 percent of the total amount of data prepared (302 instances) was used in ANN to train the network in the learning module mentioned in subsection 5.4.1. The training consisted of the input (*oilFactor*) data presented in Table 5.1, trained with the *price* as output. In order to anticipate the outcome from the network training, the lowest absolute error is attained. The remaining 10 percent of the total data are used in the predicting module as a 'blank canvas' to predict the outcome based on the remaining input of the total *oilFactor* data and to test the network's generalisation. Each ratio is individually trained with the three different activation functions mentioned in equation 5.4 and with two different values of neurons for the hidden layer, as stated in Table 5.2, in order to analyse the sensitivity of each data set.

Normalised linguistic data (the data represented in the range of -1 to 1) were used as the standard data representation in the linguistic prediction model so as to minimise noise and to relieve the networks of extensive calculations. The normalised linguistic data were used to train the learning module; thus, the sets of data, namely directional prices, normalised discrete prices and original prices were compared for accuracy. Directional prices were used as one of the training sets to map, explore and anticipate the movement of the price market according to the linguistic information gained from Chapter 4. The original prices are used in the linguistic prediction model to compare its usability in adding value to a conservative numeric prediction model as a good form of data representation for the linguistic prediction model.

These outputs were encoded to produce predictions based on the testing data provided in the ratios, in order to check the generalisation and they acted as the 'blank canvases' for the network. When evaluating the results obtained from this sensitivity analysis, root mean squared error (RMSE), normalised mean squared error (NMSE) and directional statistics (D_{stat}) were employed as a means of performing a sensitivity analysis. The summary of the results obtained from these simulations is recorded in Tables 5.3 to 5.9 of this chapter. The experiment computations and results are documented in the Appendix C.

5.4.2.1 Sensitivity Analysis: Results from Normalised Prices as the Training Output with Four and Five Neurons in the Hidden Layer

Directional price is an input obtained as one of the *oilFactor* variables obtained by sentiment mining in the fragments of articles justified in section 4.5 of Chapter 4. It was justified that the *oilFactor* input obtained from the linguistic data mined in the Google News articles contained activity connections between the inputs that explained the market's behaviour. These connections are explained through the satisfactory prediction results obtained from the sensitivity analysis. The sensitivity analysis was done by processing 24 different training sets according to the ratios mentioned in subsection 5.4.2, with two different hidden neurons and three different activation functions: the zero-based log sigmoid, the log sigmoid and the hyperbolic tangent. The results from this simulation are recorded in Tables 5.3, 5.4 and 5.5 of this sub-section. With this directional price, over half of the results showed more than 80% correct directional accuracy, while two ratios were 100% correct. The results are indicated by the directional statistic (D_{stat}) as the main performance indicator for this type of price.

Table 5.3 The Linguistic Prediction Model Results with Training Output= Directional Price, Hidden Neuron= 4.

Training: testing Ratio (%)	**Zero-based Log Sigmoid**		**Log Sigmoid**		**Hyperbolic Tangent**	
	RMSE	**Dstat (%)**	**RMSE**	**Dstat (%)**	**RMSE**	**Dstat (%)**
90:10	1.997	50.000	1.649	50.000	1.913	50.000
80:20	1.331	100.000	1.127	80.000	0.081	80.000
70:30	1.432	71.430	1.432	71.430	1.336	71.430
60:40	1.497	80.000	1.586	80.000	1.497	80.000

Table 5.4 The Linguistic Prediction Model Results with Training Output= Directional Price, Hidden Neuron= 5.

Training: testing Ratio (%)	Zero-based Log Sigmoid		Log Sigmoid		Hyperbolic Tangent	
	RMSE	Dstat (%)	RMSE	Dstat (%)	RMSE	Dstat (%)
90:10	1.568	50.000	1.806	50.000	1.665	100.000
80:20	1.314	80.000	1.111	80.000	1.082	80.000
70:30	1.433	71.430	1.422	71.430	1.332	71.430
60:40	1.521	80.000	1.584	80.000	1.533	80.000

The best directional price prediction outcome was obtained from the 80:20 percent ratio with the use of four hidden neurons and the zero-based log sigmoid as the activation function. The use of these parameters achieved results of 100% $D_{stat,}$ and 1.331 root mean squared error (RMSE) value, as depicted in Table 5.5. Directional accuracy was chosen as the superior performance indicator in this directional price simulation, as it indicates and anticipates the direction of the price market more successfully. Nevertheless, by referring to Table 5.5, it can be seen that the result derived from the second parameter is more realistic, with a RMSE value of 0.081 and a D_{stat} value of 80%. A lower RMSE value will normally reflect better prediction accuracy. Therefore, the simulation ran with the parameters of four hidden neurons and the hyperbolic tangent as the activation function, while the 80:20 percent training: testing ratio was selected as the best performance for the directional price simulation.

Table 5.5 The Best Linguistic Prediction Model Results for Training Output= Directional Price with Hidden Neuron= 4 and 5.

HIDDEN NEURON	TRAINING: TESTING RATIO (%)	ACTIVATION FUNCTION	PERFORMANCE INDICATOR	
			RMSE	Dstat (%)
4	80:20	Zero-based Log Sigmoid	1.331	100.000
4	80:20	Hyperbolic Tangent	0.081	80.000
5	90:10	Hyperbolic Tangent	1.665	100.000

5.4.2.2 Sensitivity Analysis: Results from Normalised Prices as the Training Output with four and five Neurons in the Hidden Layer

To evaluate the performance of these results, the normalised mean squared error (NMSE) was employed as the main performance indicator, with the directional statistic (D_{stat}) as the subordinate indicator. The normalised price, which minimises noise, is a reliable form of pre-processed data for training the network, as is proved in the results presented in Table 3.11 in Chapter 3. In order to evaluate the best result from this price, the lowest NMSE values were used to decide which activation function would be utilised. The best result from the normalised price simulation (Tables 5.6 and 5.7) was chosen based on the comparison of the results obtained from simulation made with four and five hidden neurons. From the observations made in Tables 5.6 and 5.7, the lowest NMSE was obtained from the training set with a 60:40 per cent ratio for the network with four hidden neurons in Table 5.6, and for the 80:20 per cent ratio in the network with five hidden neurons as shown in Table 5.7. The former ratio provides a lower NMSE, but it has poor D_{stat} values compared to the latter. The subordinate indicator was employed to validate the performances. The best result obtained in Table 5.6 was from the training set with the 90:10 per cent ratio and the log sigmoid activation function with the results of 0.685 NMSE and 100% D_{stat}.

Table 5.6 Linguistic Prediction Model Results with Training Output= Normalised Price, Hidden Neuron= 4.

Training: testing Ratio (%)	**Zero-based Log Sigmoid**		**Log Sigmoid**		**Hyperbolic Tangent**	
	NMSE	**Dstat (%)**	**NMSE**	**Dstat (%)**	**NMSE**	**Dstat (%)**
90:10	0.775	100.000	0.685	100.000	0.866	100.000
80:20	0.838	60.000	0.884	60.000	0.871	60.000
70:30	1.126	71.430	1.248	43.000	1.127	42.860
60:40	0.395	60.000	0.399	60.000	0.392	50.000

Table 5.7 Linguistic Prediction Model Results with Training Output= Normalised Price, Hidden Neuron= 5.

Training: testing Ratio (%)	**Zero-based Log Sigmoid**		**Log Sigmoid**		**Hyperbolic Tangent**	
	NMSE	**D*stat* (%)**	**NMSE**	**D*stat* (%)**	**NMSE**	**D*stat* (%)**
90:10	1.357	50.000	0.774	50.000	0.716	50.000
80:20	0.977	60.000	0.832	60.000	0.691	80.000
70:30	1.184	42.860	1.274	43.000	0.964	57.140
60:40	1.039	100.000	1.284	100.000	1.254	100.000

The results from both simulations were compared and the best result for the normalised price output was gained from the hyperbolic tangent function with 80% D_{stat} and an acceptable error in predicting (NMSE) value of 0.691, as stated in Table 5.8. Nevertheless, since D_{stat} is later used as the subordinate and the main indicator for finding the best directional accuracy with the lowest NMSE, the log sigmoid function was selected in order to offer better results for both D_{stat} and NMSE indicators, at 100% D_{stat} with 0.685 NMSE. Thus, the lowest NMSE and the highest D_{stat} were derived from the network with the log sigmoid activation function, four hidden neurons and a 90:10 per cent ratio.

Table 5.8 The Best Linguistic Prediction Model Results for Training Output= Normalised Price with Hidden Neuron= 4 and 5.

HIDDEN NEURON	**TRAINING: TESTING RATIO (%)**	**ACTIVATION FUNCTION**	**PERFORMANCE INDICATOR**	
			NMSE	**D*stat* (%)**
4	90:10	Log Sigmoid	0.685	100.000
5	80:20	Hyperbolic Tangent	0.691	80.000

5.4.2.3 Sensitivity Analysis: Results from the Original Prices as the Training Output with Four and Five Neurons in the Hidden Layer

The original price was used as the training output in order to provide a comparison and to check its suitability as a data representation for this prediction model. The suitability of this data representation was assessed based on the absolute errors that it generated from the network training. Although poor results were obtained in Chapter 3 for the use of this form as data representation (Table 3.8 shows examples of these results), employing linguistic data as well would probably increase the accuracy of the previous quantitative prediction model. To evaluate this, the root mean squared error (RMSE) was employed as the primary indicator in order to determine if the data produced the lowest error available with directional statistics ($D_{stat)}$ as the subordinate. Poor results were obtained from this form of price, with most of the RMSE values attained being more than 20.000, as presented in Tables 5.9 and 5.10. This poor RMSE result is also shown in the subordinate indicator through poor D_{stat} values of less than 50%, with most showing 0% the best RMSE values. This is believed to be caused by the data being presented in time-series form and not being standardised to the range between -1 to 1 (normalised).

Table 5.9 Linguistic Prediction Results with
Training Output= Original Price, Hidden Neuron= 4.

Training: testing Ratio (%)	**Zero-based Log Sigmoid**		**Log Sigmoid**		**Hyperbolic Tangent**	
	RMSE	**D*stat* (%)**	**RMSE**	**D*stat* (%)**	**RMSE**	**D*stat* (%)**
90:10	29.662	50.000	22.673	0.000	33.345	0.000
80:20	45.482	20.000	40.211	20.000	34.481	20.000
70:30	32.893	42.860	53.463	42.850	35.919	28.57
60:40	42.052	60.000	44.606	50.000	31.261	50.000

Table 5.10 Linguistic Prediction Results with
Training Output=Original Price, Hidden Neuron= 5.

Training: testing Ratio (%)	Zero-based Log Sigmoid		Log Sigmoid		Hyperbolic Tangent	
	RMSE	D*stat* (%)	RMSE	D*stat* (%)	RMSE	D*stat* (%)
90:10	29.138	0.000	26.276	0.000	42.435	0.000
80:20	43.337	20.000	41.837	20.000	30.705	40.000
70:30	35.747	42.860	51.253	42.860	37.398	28.570
60:40	41.084	40.000	46.655	50.000	37.130	60.000

Training the normalised inputs with this time-series output data probably caused difficulties for the network training. Nevertheless, when comparing the data results with the respective activation output functions, both types of hyperbolic tangent occasionally deliver the lowest RMSE for both hidden neurons used, as evidenced in Table 5.11.

Table 5.11 The Best Linguistic Prediction Model Results for Training Output= Original Price with Hidden Neuron= 4 and 5.

HIDDEN NEURON	TRAINING: TESTING RATIO (%)	ACTIVATION FUNCTION	PERFORMANCE INDICATOR	
			RMSE	D*stat* (%)
4	60:40	Hyperbolic Tangent	31.261	50.000
5	90:10	Log Sigmoid	26.276	0.000

5.4.2.4 Overall Result Evaluation for the Linguistic Prediction Model

Observations were made in sections 5.4.2.1 to 5.4.2.3 to consider, select and compare the values of the parameters that produced the optimal performance for each price prediction in these sections. The best performances were considered based on the lowest root mean squared error (RMSE) and the normalised mean squared error (NMSE) value with the highest directional statistics (D_{stat}) value from each of the directional, normalised and original price forms. Nonetheless, although the best results were chosen from each price model, the best of the three was selected and thus represents the overall best performance of the model. The final results represent the overall best performance of the linguistic prediction model and are presented in Table 5.12. From the observations made during all the experiments for the linguistic prediction model, it is explained that the use of the hyperbolic tangent function as an accelerator enhanced the network and contributed to better prediction results. This is shown by the promising results produced by the prediction network via the use of the hyperbolic tangent as the activation function, as evidenced in Tables 5.5, 5.8, and 5.11. The hyperbolic tangent activation function provided the most accurate results of all three performance indicators, over all training ratios and all price models in the experiment.

Table 5.12 The Best Simulation Results for Linguistic Prediction Model.

PRICE	HIDDEN NEURON	TRAINING: TESTING RATIO	ACTIVATION FUNCTION	PERFORMANCE INDICATOR		
				RMSE	NMSE	D*stat*%
Directional	4	80:20	Hyperbolic Tangent	0.081	-[173]	80.000
Normalised	4	90:10	Log Sigmoid	-	0.685	100.000
Original	5	90:10	Log Sigmoid	26.276	-	0.000

[173] "-" is given to the performance indicator that is not considered to be the primary or subordinate indicator.

5.4.2.5 Comparison of the Linguistic Prediction Model Results with other Machine Learning Approaches

In order to measure the performance of our linguistic back-propagation neural network (BPNN), a comparison of this linguistic prediction model was made with the other machine learning methods presented in Chapter 2 of subsection 2.1.2. The support vector machine (SVM) with support vector regression (SMO) [108] was used for comparison with the linguistic prediction model by implementing polykernel as its kernel and improved SMO regression as its regression optimiser property. The same data ratios of 90:10 percent, 80:20 percent, 70:30 percent and 60:40 percent training: testing was used to compare the result obtained from the linguistic prediction model with all other methods. A linear regression (LR) with the Akaike criterion method assists model selection was also applied for this purpose. In addition, a Gaussian process (GP) with an RBF kernel was also chosen for this comparison. The experiments made for this comparison were conducted via an open-source programme called Weka [109]. The performance indicators implemented for these experiments are based on the values of the root mean squared error (RMSE) and the mean absolute error (MAE) obtained from each ratio. The MAE is computed by the formula in equation 5.10.

$$MAE = \frac{1}{n}\sum_{i=1}(e_i) = \frac{1}{n}\sum_{i=1}(\hat{x}_i - x_i) \tag{5.10}$$

where variables x_i and $\hat{x}_i$ denote the actual and predicted values, respectively. Meanwhile, e_i denotes the error obtained from the difference attained in the prediction. From the experiments conducted, it was discovered that the BPPN offers a fairly good result compared with the other methods observed. The observations began by examining the best RMSE and MAE from each method in each ratio in Table 5.13. The objective of the observation was to find the best and the second best result for the comparison by selecting the method with the lowest value for each ratio. Through these observations, it was found that the GP had the best result, with 75% of its results having the smallest RMSE and MAE values compared with the other two approaches. This was followed by SVM and LR. Nonetheless, the smallest RMSE and MAE values of 0.9861 and 0.9081, respectively, were discovered in the

GP's 80:20 ratio and in the SVM's 80:20 ratio. Compared with our linguistic prediction model in subsection 5.4.2.4, the results were evaluated based on the RMSE and the NMSE.

Throughout the observation, the BPNN achieved better performance in terms of its lowest values for both the RMSE and the NMSE. Although the LR offers zero NMSE value, its RMSE value is bigger than that of the BPNN. In order to train the data with the LR, it was set to select the data with the M5 method, where the data with the smallest standardised coefficient (number) were removed until no improvement was observed in the estimated error. The value of the LR's ridge parameter was set to 1.0E-8. Since the data used in this prediction model were represented in the normalised form, the NMSE will be the superior indicator in evaluating the linguistic prediction model performance in comparison with the other machine learning approaches mentioned in this subsection. The NMSE was also utilised as the primary indicator in subsection 5.4.2.2 of this chapter.

Kumar, et al. [110] suggested that, in order for a model to be reliable, and to be acceptable, the NMSE value should be 0.5 or less. Nevertheless, the NMSE was used as the performance indicator in this linguistic prediction model, as it measures the total error of a prediction model [111]. A model is deemed to be good if it produced the minimum NMSE value compared to the other models. In descending order, the next best prediction methods after the BPNN, are the LR, the SVM and the GP. These prediction results are presented in Tables 5.13 and 5.14.

Table 5.13 The Results of the Linguistic Prediction Model from The Support Vector Machine (SVM), the Linear Regressions (LR) and the Gaussian Process (GP) Approaches.

TRAINING: TESTING RATIO	SVM		LR		GP	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
90:10	1.5909	1.5795	1.6102	1.6041	1.2327	1.2157
80:20	1.0001	0.9081	1.0592	0.9138	0.9861	0.9469
70:30	2.072	1.7242	2.5221	2.2572	1.0229	0.9975
60:40	1.9135	1.5481	1.0000	1.0000	1.0023	0.9918

Table 5.14 The Comparison of the Best Results from the Support Vector Machine (SVM), the Linear Regressions (LR) and the Gaussian Process (GP) Approaches with the Back-Propagation Neural Networks (BPNN) used in the Linguistic Prediction Model.

BPNN		SVM		LR		GP	
RMSE	NMSE	RMSE	NMSE	RMSE	NMSE	RMSE	NMSE
0.081	**0.685**	1.000	11.105	1.0000	0.0000	0.9861	24.8304

5.5 THE LINGUISTIC AND QUANTITATIVE (LQ) PREDICTION MODEL: THE HYBRID

Previously, the hybridisation of the linguistic and quantitative data and the model used has proven to provide interesting results for some specific domains, according to the literature review in section 5.1. We hypothesised earlier, in section 5.1, that this combination should somehow show similar pattern trending as the domains discussed in the previous literature mentioned in section 5.1 of this chapter. Combining the linguistic data with a quantitative prediction model adds value to the prediction model by enhancing its performance. In this chapter, the discussion in section 5.5.1 is presented in Figure 5.4, which illustrates the development of our final prediction model — the Linguistic-Quantitative (LQ) model. The LQ model was developed to explore the probability of this hybridisation being able to encode both linguistic and quantitative data in the hybrid prediction model. The model is designed to combine two different forms of data into a single prediction model.

From the quantitative prediction model discussed in section 3.3 of Chapter 3 and the linguistic prediction model discussed in section 5.4 of this chapter, it was seen that both of these individual models, the quantitative prediction model and the linguistic prediction model, perform competitively separately from each other. We believe that integrating these models will improve prediction performance. During the model development process, we examined the contribution of the linguistic prediction model and the quantitative prediction model to the task of extracting knowledge and integrating it into the LQ model. The LQ prediction model development is discussed further in the next sections of this chapter.

5.5.1 Back-Propagation Neural Network (BPNN) Application in the Linguistic-Quantitative (LQ) Hybrid Prediction Model

ANN have been utilised in both linguistic and quantitative prediction models and have proven to be a promising prediction method for the market. In order to extend the predictive models developed so far, we employed another ANN in this hybrid model to integrate both models and to validate its complementary ability for predictions in the research domain market. Figure 5.4 presents the development framework constructed to integrate these two prediction models.

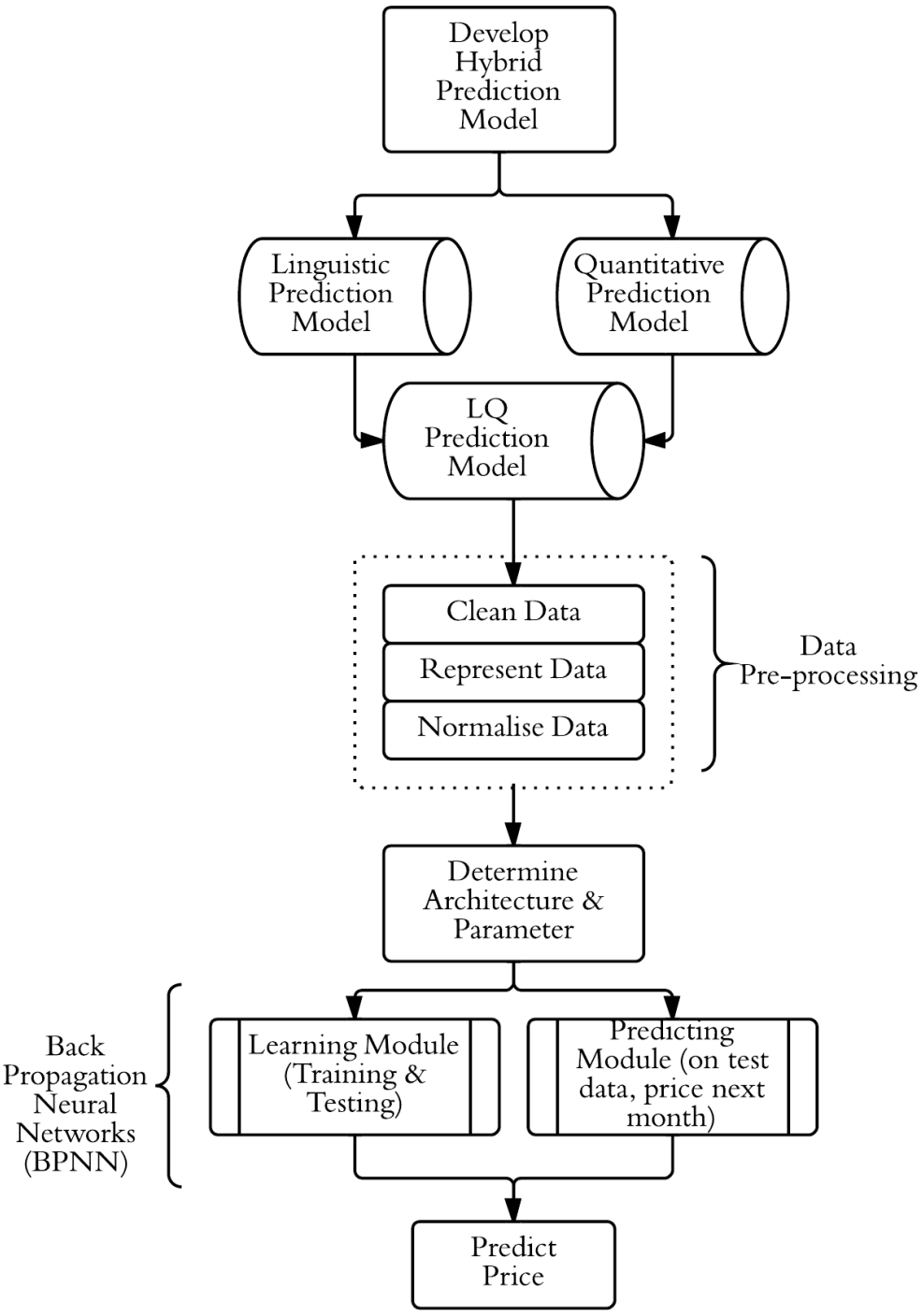


Figure 5.4 The Linguistic-Quantitative (LQ) Prediction Model Framework with the Back-Propagation Neural Network (BPNN).

The development of the linguistic and quantitative (LQ) prediction model begins with the data pre-processing phase, where data in the linguistic and quantitative model are normalised in order to reduce noise and to provide a standard input for the ANN. The normalised data are then used to feed this model as a continuation of the standardisation[174] made in the previous two prediction models. Before deciding on the network architecture, the data are first sub-divided into a number of training sets in order to train the ANN. From the discussions made in section 3.3 and section 5.4, we have seen models that house the different dates and the different range of data and which are therefore different kinds of input. Thus, the standardisation of these data is required in order to provide a standard data set for training.

The input data for the hybrid model are made up of both linguistic and quantitative data. The quantitative data were derived from the *oilFactor* features used in the quantitative prediction model, as discussed in section 3.3. The linguistic data used for this LQ model were derived from the linguistic features discussed in sections 5.2 to 5.4 and are presented in Table 5.1.

The training in the LQ prediction model was divided into three different sets of data, based on three different time durations. A full range of monthly quantitative data dated from January 1984 to December 2009 was used in training set A, coupled with the monthly linguistic data dated from January 2008 to December 2009. Next, a range of monthly quantitative data dated from January 2006 to December 2009 was used in training set B, coupled with the same monthly linguistic data employed in training set A. Meanwhile, training set C consisted of monthly quantitative data dated from January 2008 to December 2009 and was coupled with the monthly linguistic data ranging from January 2008 to December 2009. These training divisions were designed to check the BPNN's capability to handle training with incomplete data and training with a small data matrix. Also, these sets of data were depicted into the network in order to examine the linguistic data's credibility in igniting the connections between the inputs in the network, especially when combined with quantitative data. This analysis is discussed further through the sensitivity analysis discussions in subsection 5.6.1 of this chapter.

[174] The data was normalised, based on equation (3.1), to contain numbers ranging from {-1 to 1}.

The training sets that will be used in the LQ model are presented in Table 5.15 and have the following elements:

i) Training set A: a time period with complete quantitative data, but with incomplete linguistic data;
ii) Training set B: a time period with complete quantitative data, but with linguistic data that is more complete; and
iii) Training set C: a time period with complete quantitative and linguistic data.

Table 5.15 Data Sets for the Linguistic-Quantitative (LQ) Prediction Model.

SET	DATA RANGE	DATA FORM	CONTENT	TOTAL DATA
A	Jan ’84-Dec ‘09	Normalised	Complete quantitative data. Incomplete linguistic data from Jan ’84-Dec ’07.	312 data with 14,976 observations.
B	Jan ’06-Dec ‘09	Normalised	Complete quantitative data. Less complete linguistic data. Incomplete linguistic data from Jan ’06-Dec ’07.	48 data with 2,304 observations.
C	Jan ’08-Dec’09	Normalised	Complete linguistic and quantitative data Jan ’08-Dec ’09.	24 data with 1,152 observations.

5.6 THE LINGUISTIC-QUALITATIVE (LQ) PREDICTION MODEL: ITS NETWORK ARCHITECTURE AND PARAMETERS

Combining two different models into a singular hybrid network is known to be a challenging task. To explore and use simpler methods would be a good solution to achieve the research objectives. BPNN was applied in this linguistic-quantitative (LQ) model to explore its usability for predictions when using two different types of data. The application of BPNN in this LQ model also proved that the use of BPNN as a prediction tool is still relevant and it is not obsolete. Its relevancy was proven by the promising results it generated in the quantitative prediction model and the linguistic prediction model discussed in section 3.4 of Chapter 3 and sub-section 5.4.2 of Chapter 5.

Wang, et al. [39] agreed that the utilisation of the BPNN technique as an integration tool for prediction was proven to be competitive and reliable. Hence, it was decided to utilise ANN once again as the integration tool for our models. The results gained from both the quantitative prediction model and the linguistic prediction model had shown promising outcomes that indicated BPNN's relevance for being used again as an ANN technique in the LQ model. The LQ model was developed via a similar framework that was used for the quantitative prediction model and for the linguistic prediction model. The architecture and the parameters for the network are depicted in Table 5.16.

Table 5.16 Back-Propagation Neural Network (BPNN) Parameters for the Linguistic-Quantitative (LQ) Prediction Model.

PARAMETER	VALUE
Initial Weight	0.3
Learning Rate	0.3
Momentum	0.6
Learning Cycle (epochs)	5000
Network Layer	3
Neurons:	
Input Layer	$22^{176}+13^{177} = 35$
Hidden Layer	4
	5
Output Layer	1
Sigmoid	Zero-based Log Sigmoid
	Log Sigmoid
	Hyperbolic Tangent

The architecture and parameters employed in this model are similar to those in the previous quantitative prediction model and in the linguistic prediction model discussed in section 3.4 and sub-section 5.4.2. The only difference with this LQ model is the number and the combination type of input neurons used for training. The input data that were used for the training in Chapter 3 (quantitative data) and Chapter 5 (linguistic data) were proven to generate a good interconnection of neuron activities. This good interconnection was manifested by the minimum error result produced by both the quantitative prediction model and the linguistic prediction model in section 3.4 and sub-section 5.4.2.

We believe that combining these two kinds of data (with good interconnections) as input for the LQ model will improve the interconnections and will provide better prediction outcomes. Good prediction outcomes will further establish better credibility for ANN continuing to be a reliable tool for predicting the crude oil price. Hence, the 22 input layers of the quantitative prediction model, containing quantitative features and 13 input layers from the linguistic prediction model, which were then combined as input in the network structures. A sensitivity analysis was carried out to analyse the sensitivity of the network when propagating with four and

[176] Quantitative features
[177] Linguistic features

five hidden neurons. The network topology is visualised in Figure 5.5 as a three-layer ANN.

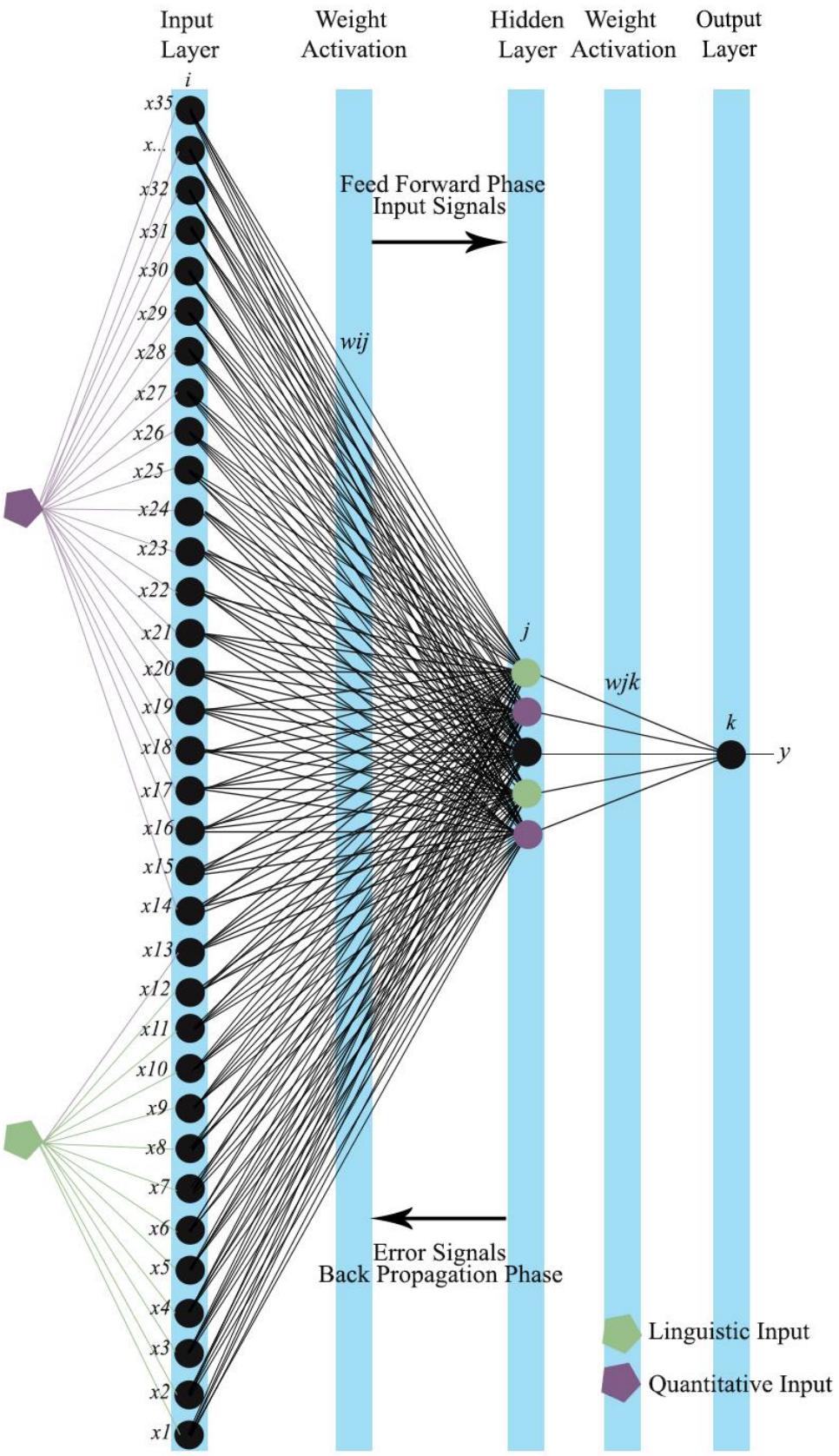


Figure 5.5 Three-Layer Artificial Neural Networks (ANN) Trained with Back-Propagation Topology of 35-5-1 for the Linguistic-Quantitative (LQ) Prediction Model

The inputs *X* used in this linguistic-quantitative (LQ) prediction model were feed-forward activated with the function *f* to propagate interconnections between the layers. The signal later back-propagated the errors that were produced in output layer *Y*. The process continues to loop until a desired output (the one with the smallest error) is produced from the network. The feed-forward phase and the back-propagation phase presented in Figure 5.5 are mathematically represented in equation 5.11:

$$
\begin{aligned}
&Y = f(X_l) \\
&\text{with } X_l = \{x_{l1}, x_{l2,} x_{l\ldots,} x_{ln}\} \text{ of linguistic inputs and} \\
&\\
&Y = f(X_q) \\
&\text{with } X_q = \{x_{q1}, x_{q2,} x_{q\ldots,} x_{qn}\} \text{ of quantitative input}
\end{aligned}
\tag{5.11}
$$

The integration of the quantitative prediction model and the linguistic prediction model as a linguistic-quantitative prediction (LQ) model are mathematically shown by equation 5.12, where *i, j* and *k* denote the input-hidden-output connection, while *w* and *e* denote the weight and error values of the final connection.

$$
y_k = f\left[\sum_{j=1}^{m}\left(x_j^l w_{jk}\right) + \left(x_j^q w_{jk}\right)\right] - e_k \tag{5.12}
$$

5.6.1 Sensitivity Analysis: Simulation Test and Results

In order to analyse the sensitivity of the network depicted in Table 5.16, three sets of training were prepared according to sub-section 5.5.1 and as summarised in Table 5.15, with three different training: testing ratios[178]. The root mean squared error (RMSE), the normalised mean squared error (NMSE) and the directional statistics (D_{stat}) were used to compare the experimental results, with NMSE as the primary indicator and D_{stat} the subordinate. The accuracy of this model is evaluated primarily according to the lowest value of NMSE: the lower the NMSE, the more accurate the model. Nevertheless, the results were also interpreted based on the accuracy of its target direction. The accuracy of a target discrete price, combined with the accuracy of a target price's direction, will add credibility to the network model.

Referring to Table 5.15, the data used in this LQ model are divided into three sets:
i) Set A consists of data ranging from January 1984 to December 2009 (25 years), with missing linguistic data from January 1984 to December 2007. An investigation was made with this data set A to evaluate the credibility of network prediction using the quantitative data combined with the small set of linguistic data.

ii) Set B consists of data ranging from January 2006 to December 2009 (3 years), with missing linguistic data for the time period of January 2006 to December 2007. The data set B was prepared in order to observe the credibility of the small set of linguistic data prepared for the network in complementing the smaller set of quantitative data to be used in the network. This also aimed to examine the contribution of a linguistic data set to supplement the network with the quantitative data. The network capability in propagating with an almost full[179] data set but with a shorter time period was also investigated.

[178] The training: testing ratios are the data cross-validation technique used for LQ model that employed 90 percent of the total data for training and 10 percent for testing (90:10), 80 percent of total data for training and 20 percent for testing (80:20), and 70 percent of total data for training and 30 percent for testing (70:30).

[179] Almost full data set: in this context, less missing data were distributed into the network compared to data set A.

iii) Set C consists of data ranging from January 2008 to December 2009 (12 months) and was designed to investigate the network's capability for prediction with well-distributed data (full data without missing values), but with a very short time period for learning.

The values of RMSE and NMSE were determined based on the minimum value post-training from each set {A, B, C}. The primary performance indicator for these experiments was NMSE, followed by RMSE and D_{stat} as the subordinates. It should be noted that RMSE and NMSE values were determined as follows: the closer the RMSE and NMSE values were to zero, the more accurate the outcome. With regard to D_{stat}, the optimal values are determined according to the highest percentage D_{stat} that the sets produced. The performances of these sets are evaluated based on the following rules:

1. From set {A, B, C}, choose the lowest NMSE value from each training: testing ratio:
- Choose the lowest NMSE from set A|B|C-1[180].
- Choose the lowest NMSE from Set A|B|C-2[181].
- Compare and highlight the minimum value from the above two.

2. From set {A, B, C}, choose the highest D_{stat} value from each training: testing ratio:
- Choose the highest D_{stat} from Set A|B|C-1.
- Choose highest D_{stat} from Set A|B|C-2.
- Compare and highlight the maximum value from the above two.

3. From set {A, B, C}, choose the lowest RMSE value from each training: testing ratio:
- Choose the lowest RMSE from Set A|B|C-1.
- Choose the lowest RMSE from Set A|B|C-2.
- Compare and highlight the minimum value from the above two.

4. From set {A, B, C}, choose the data set with the most optimal performance, highlighted as the best performance result.
The results of this sensitivity analysis were recorded in subsections 5.6.1.1 to 5.6.1.6, where each set was divided into two further subsets, A|B|C-1 and A|B|C-2, indicating the number of hidden neuron sets utilised for the network training.

[180] The training set with hidden layer = 4
[181] The training set with hidden layer = 5

5.6.1.1 Sensitivity Analysis: Results from Training Set A with 4 and 5 Neurons in the Hidden Layer

The results from the simulation test made for training set A are presented in this section.

Table 5.17 The Results for the Linguistic-Quantitative (LQ) Prediction Model with Training Set A-1, Hidden Neuron= 4.

Training: testing Ratio	Zero-based Log Sigmoid			Log Sigmoid			Hyperbolic Tangent		
	NMSE	RMSE	D*stat* (%)	NMSE	RMSE	D*stat* (%)	NMSE	RMSE	D*stat* (%)
90:10	0.969	0.087	83.870	3.265	0.103	77.420	4.445	0.127	67.740
80:20	1.675	0.091	77.419	1.803	0.086	79.032	4.191	0.116	69.355
70:30	1.107	0.101	77.660	1.232	0.095	77.660	1.675	0.119	65.957

Table 5.18 The Results for the Linguistic-Quantitative (LQ) Prediction Model with Training Set A-2, Hidden Neuron= 5.

Training: testing Ratio	Zero-based Log Sigmoid			Log Sigmoid			Hyperbolic Tangent		
	NMSE	RMSE	D*stat* (%)	NMSE	RMSE	D*stat* (%)	NMSE	RMSE	D*stat* (%)
90:10	0.055	0.021	87.100	2.927	0.099	74.190	7.912	0.125	70.970
80:20	1.212	0.075	77.420	1.291	0.077	79.030	4.830	0.115	69.360
70:30	1.121	0.102	72.340	1.275	0.097	77.660	1.122	0.139	62.770

From the observation, the optimal result with the lowest NMSE and the highest D_{stat} value from training set A-1 was obtained from the 90:10 per cent training: testing ratio as logged in Table 5.17, with 0.9693 as the NMSE, 0.0867 as the RMSE and 83.870% as the D_{stat}. For training set A-2 (Table 5.18), the best performance result was also derived from the 90:10 per cent training: testing ratio, with 0.055 NMSE, 0.021 RMSE and 87.10% as the D_{stat}. These performance results were determined based on the minimum NMSE value gained from the simulation that acted as the primary performance indicator and the RMSE with D_{stat} as the subordinate indicator.

The best performances of the two subsets were recorded in Table 5.19 and represented in Figures 5.6 and 5.7 of this section.

Table 5.19 The Best Results for the Linguistic-Quantitative (LQ) Prediction Model for Training Set A.

SET	TRAINING: TESTING RATIO	HIDDEN NEURONS	ACTIVATION FUNCTION	PERFORMANCE INDICATOR		
				NMSE	RMSE	D*stat* (%)
A-1	90:10	4	Zero-based Log Sigmoid	0.969	0.087	83.870
A-2	90:10	5	Zero-based Log Sigmoid	0.055	0.021	87.100

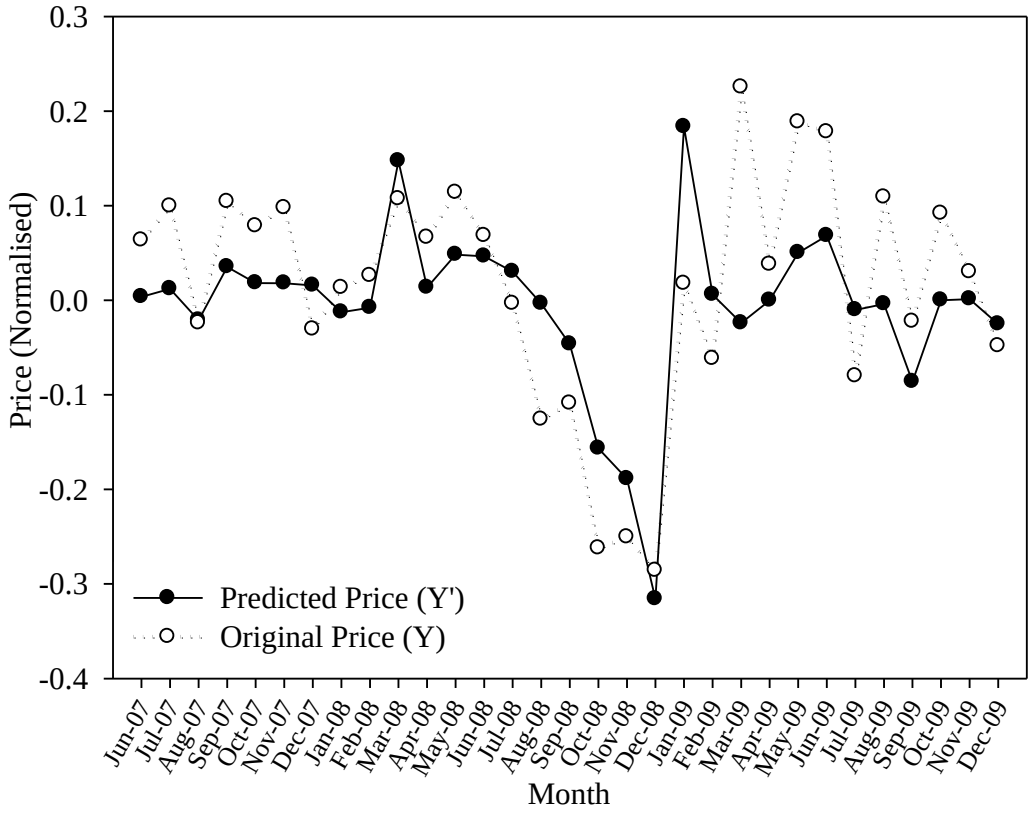


Figure 5.6 The Best Performance Result for Training Set A, Subset A-1 of the Linguistic-Quantitative (LQ) Prediction Model with 90:10 Per cent Data Ratio and 4 Hidden Neurons.

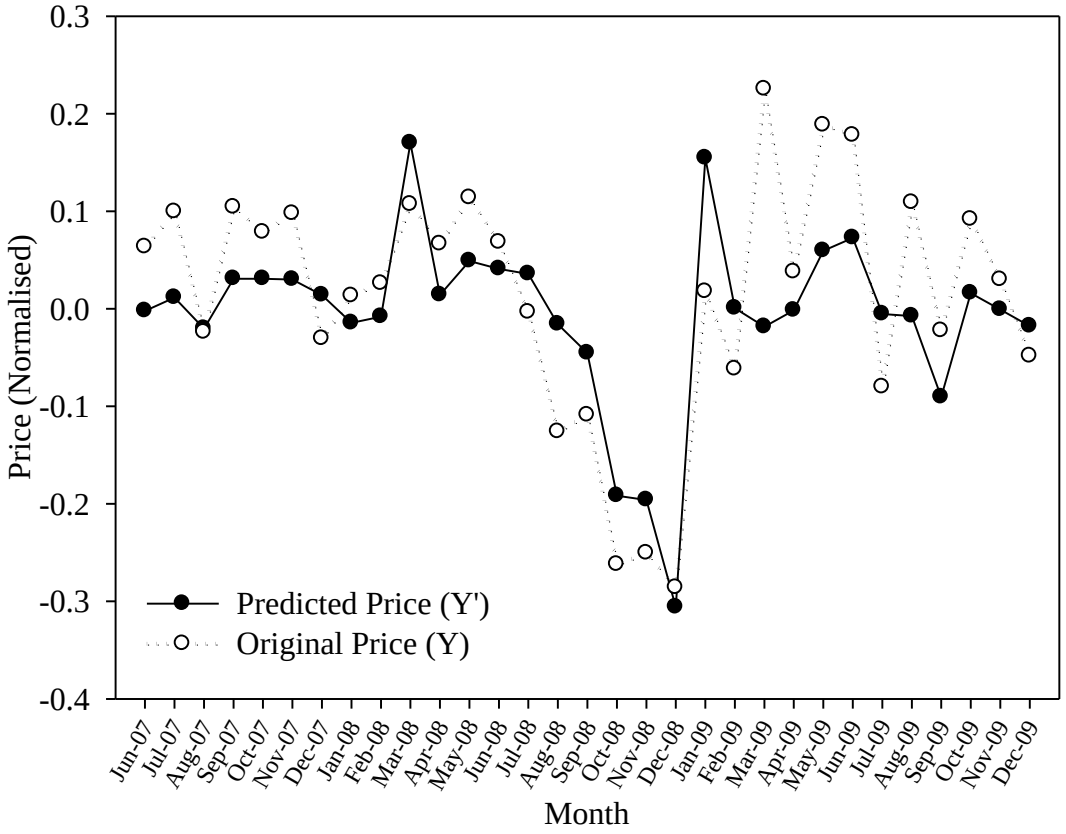


Figure 5.7 The Best Performance Result for Training Set A, Subset A-2 of the Linguistic-Quantitative (LQ) Prediction Model with 90:10 Per cent Data Ratio and 5 Hidden Neurons.

Figures 5.6 and 5.7 show that the directional performance for these sets is good. The figures show maximum positive and parallel movements of predicted price against the movements of the original price. The NMSE and the RMSE values obtained from this sensitivity simulation also show promising results. Overall, the results gained from this training set are satisfactory, as the percentage of linguistic data combined with the quantitative input is small.

5.6.1.2 Sensitivity Analysis: Results from Training Set B with 4 and 5 Neurons in the Hidden Layer

In this section, the training involved a lesser amount of quantitative data and linguistic data that was less incomplete than that of set A. The results from the sensitivity analysis of Set B were presented in Tables 5.20 to 5.22 of this section. The best performance in set B was chosen according to the minimum NMSE value and the maximum D_{stat} value as the primary and secondary performance indicators.

Table 5.20 The Results for the Linguistic-Quantitative (LQ) Prediction Model with Training Set B-1, Hidden Neuron= 4.

Training: testing Ratio	**Zero-based Log Sigmoid**			**Log Sigmoid**			**Hyperbolic Tangent**		
	NMSE	**RMSE**	**D*stat* (%)**	**NMSE**	**RMSE**	**D*stat* (%)**	**NMSE**	**RMSE**	**D*stat* (%)**
90:10	1.124	0.246	40.000	1.195	0.145	60.000	0.950	0.177	60.000
80:20	1.879	0.156	90.000	2.069	0.130	90.000	2.844	0.161	80.000
70:30	1.701	0.166	64.290	1.512	0.156	71.430	1.295	0.121	85.710

Table 5.21 The Results for the Linguistic-Quantitative (LQ) Prediction Model with Training Set B-2, Hidden Neuron= 5.

Training: testing Ratio	**Zero-based Log Sigmoid**			**Log Sigmoid**			**Hyperbolic Tangent**		
	NMSE	**RMSE**	**D*stat* (%)**	**NMSE**	**RMSE**	**D*stat* (%)**	**NMSE**	**RMSE**	**D*stat* (%)**
90:10	1.406	0.245	40.000	1.049	0.155	60.000	0.960	0.191	40.000
80:20	1.632	0.183	70.000	1.934	0.128	90.000	2.452	0.186	70.000
70:30	1.948	0.183	78.570	1.578	0.150	71.430	0.426	0.181	71.430

Training set B obtained good directional performance results from both set B1 and B2, where the maximum D_{stat} values of these two sets were 90.00% accurate, as recorded in Table 5.20 and Table 5.21, respectively. The maximum D_{stat} values (90.00%) were obtained from the training with an 80:20 per cent data ratio and a

zero-based log sigmoid as its activation function in Table 5.20, and from the training with an 80:20 data ratio and a log sigmoid as its activation function in Table 5.21. This directional performance was mapped in Figures 5.8 and 5.9.

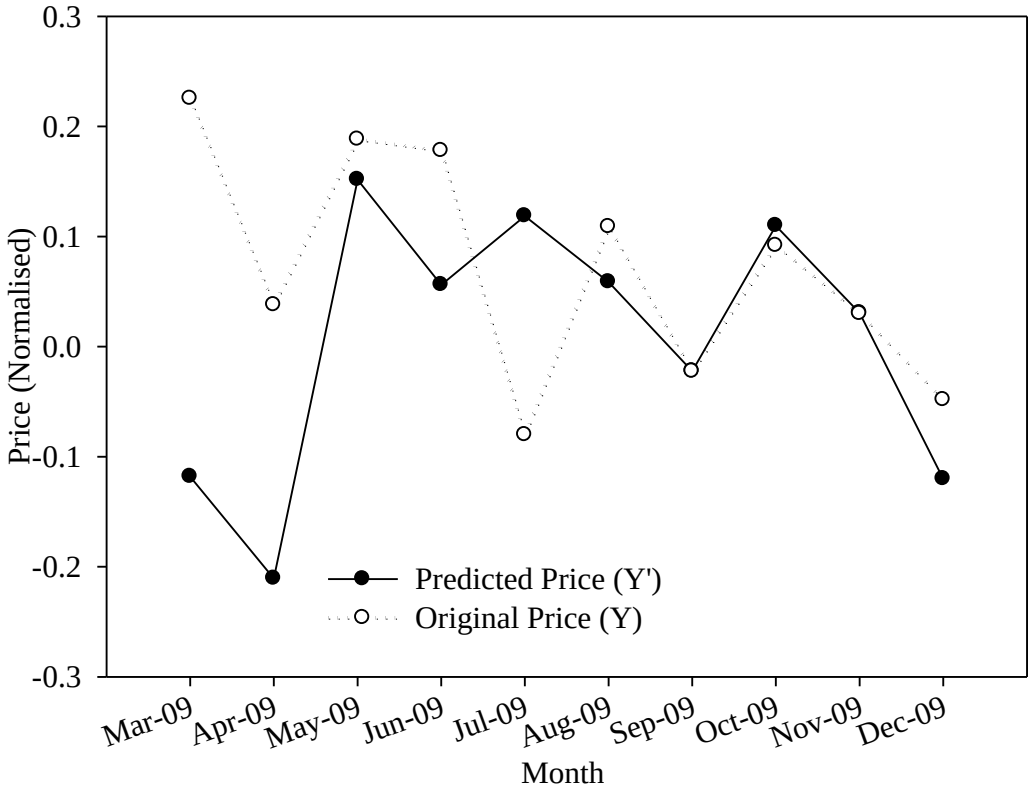


Figure 5. 8 The Best Directional Performance Result (90.00%) for Training Set B, Subset B-1 of the Linguistic-Quantitative (LQ) Prediction Model based on the Directional Statistics (D_{stat}) Value, Trained with an 80:20 Per cent Data Ratio and 4 Hidden Neurons.

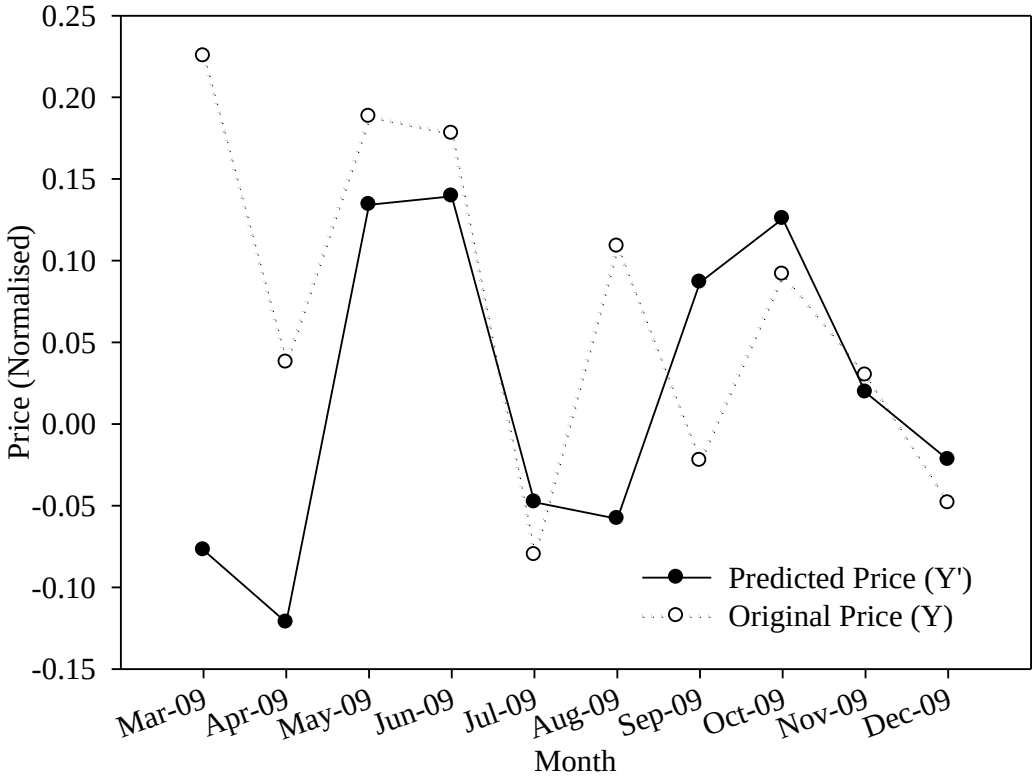


Figure 5.9 The Best Directional Performance Result (90.00%) for Training Set B, Subset B-2 of the Linguistic-Quantitative (LQ) Prediction Model based on the Directional Statistics (D_{stat}) Value, Trained with an 80:20 Per cent Data Ratio and 5 Hidden Neurons.

Although training sets B1 and B2 obtained good directional performance results, their NMSE values, 1.879 and 1.934, respectively, did not echo this because lower NMSE values would be a better performance indicator. Therefore, in order to choose the best performance for training set B, the NMSE was first used as the primary indicator. The lowest NMSE value will indicate better accuracy results for the discrete price prediction for this LQ model.

Based on Tables 5.20 and 5.21, the minimum NMSE value obtained from set B-1 was 0.950 and 0.426 for set B2, with fairly good D_{stat} values of 60.00% and 71.43%, respectively. These best performance results, based on the NMSE value as the primary indicator, are recorded in Table 5.22. Comparing these D_{stat} values with the D_{stat} values obtained in training set A (Table 5.19) reveals that the duration of the

historical data used for training the network has a significant effect on the network's directional result. Nonetheless, the values of NMSE and RMSE obtained from set B, when compared with set A, were promising. These promising values show that linguistic elements are significant as a complementary value to the quantitative prediction model, even with a smaller range of data.

Table 5.22 The Best NMSE Results for the Linguistic-Quantitative (LQ) Prediction Model for Training Set B.

SET	TRAINING: TESTING RATIO	HIDDEN NEURONS	ACTIVATION FUNCTION	PERFORMANCE INDICATOR		
				NMSE	RMSE	D*stat* (%)
B-1	90:10	4	Hyperbolic Tangent	0.950	0.177	60.000
B-2	70:30	5	Hyperbolic Tangent	0.426	0.181	71.430

The best performance of training set B is visualised in the function mapping in Figure 5.10.

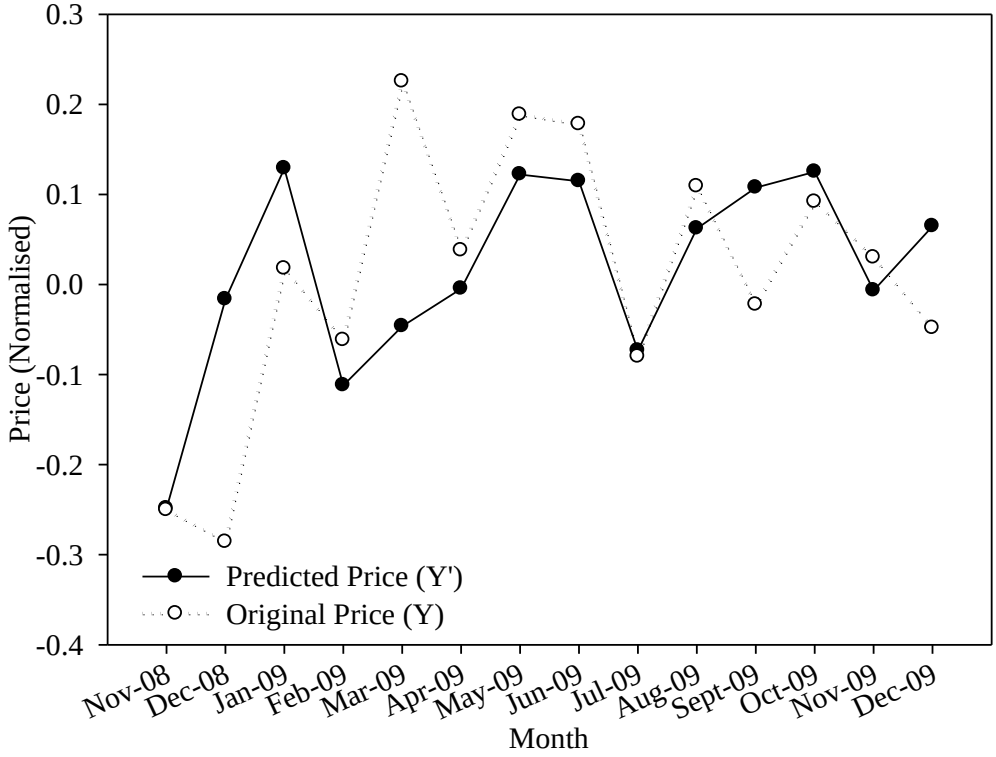


Figure 5.10 The Best Performance Result for Training Set B of the Linguistic-Quantitative (LQ) Prediction Model with a 70:30 Per cent Data Ratio and 5 Hidden Neurons.

5.6.1.3 Sensitivity Analysis: Results from Training Set C with 4 and 5 Neurons in the Hidden Layer

In this section, the trainings were conducted using a smaller amount of quantitative and linguistic data than were used in sets A and B. The linguistic inputs are equally distributed throughout the time period of January 2008 to December 2009. The results from the simulation test made for the training of Set C are presented in Tables 5.23 and 5.24.

Table 5.23 The Results for the Linguistic-Quantitative (LQ) Prediction Model with Training Set C-1, Hidden Neuron= 4.

Training: testing Ratio	**Zero-based Log Sigmoid**			**Log Sigmoid**			**Hyperbolic Tangent**		
	NMSE	**RMSE**	**D*stat* (%)**	**NMSE**	**RMSE**	**D*stat* (%)**	**NMSE**	**RMSE**	**D*stat* (%)**
90:10	0.668	0.172	50.00	0.669	0.173	50.00	0.678	0.181	50.00
80:20	0.934	0.162	80.00	1.283	0.151	60.00	1.252	0.148	60.00
70:30	0.922	0.152	71.43	0.757	0.138	71.43	0.969	0.156	85.71

Table 5.24 The Results for the Linguistic-Quantitative (LQ) Prediction Model with Training Set C-2, Hidden Neuron= 5.

Training: testing Ratio	**Zero-based Log Sigmoid**			**Log Sigmoid**			**Hyperbolic Tangent**		
	NMSE	**RMSE**	**D*stat* (%)**	**NMSE**	**RMSE**	**D*stat* (%)**	**NMSE**	**RMSE**	**D*stat* (%)**
90:10	0.684	0.186	50.00	0.668	0.171	50.00	0.652	0.164	50.00
80:20	0.938	0.139	80.00	1.299	0.148	60.00	1.467	0.169	40.00
70:30	0.682	0.151	71.43	0.739	0.134	71.43	0.920	0.164	71.43

Similar to sets A and B, the primary performance indicator for set C is the minimum value of NMSE, and the maximum value of D_{stat} is the secondary performance indicator. From the observations in Tables 5.23 and 5.24, the minimum NMSE values (0.668) for the training of set C-1 were obtained from the training set with a 90:10 per cent data ratio and a zero-based log sigmoid activation function. Set C-2 (0.652), which was obtained from the training set with a 90:10 per cent data ratio and a hyperbolic tangent activation function, produced minimum values of D_{stat} that were interpreted as a low level of directional performance accuracy. As both the minimum NMSE results were produced by 10% of the testing data, this is perceived to have contributed to the minimum D_{stat} values. Based on the sensitivity analysis made in sections 5.4 to 5.6, a greater percentage of testing data was needed for improved directional accuracy. Due to the lower values of D_{stat} gained from the minimum NMSE values in Table 5.23 and Table 5.24, it was decided that D_{stat} should be the primary performance indicator instead of NMSE, so as to optimise the performance result in training set C. The best performance results for training set C are presented in Table 5.25 and mapped in Figures 5.11 and 5.12, with error mapping illustrated by Figures 5.13 and 5.14.

Table 5.25 The Best Performance Results for the Linguistic-Quantitative (LQ) Prediction Model for Training Set C.

SET	TRAINING: TESTING RATIO	HIDDEN NEURONS	ACTIVATION FUNCTION	PERFORMANCE INDICATOR		
				NMSE	RMSE	D*stat* (%)
C-1	70:30	4	Hyperbolic Tangent	0.969	0.156	85.71
C-2	80:20	5	Zero-based Log Sigmoid	0.938	0.139	80.00

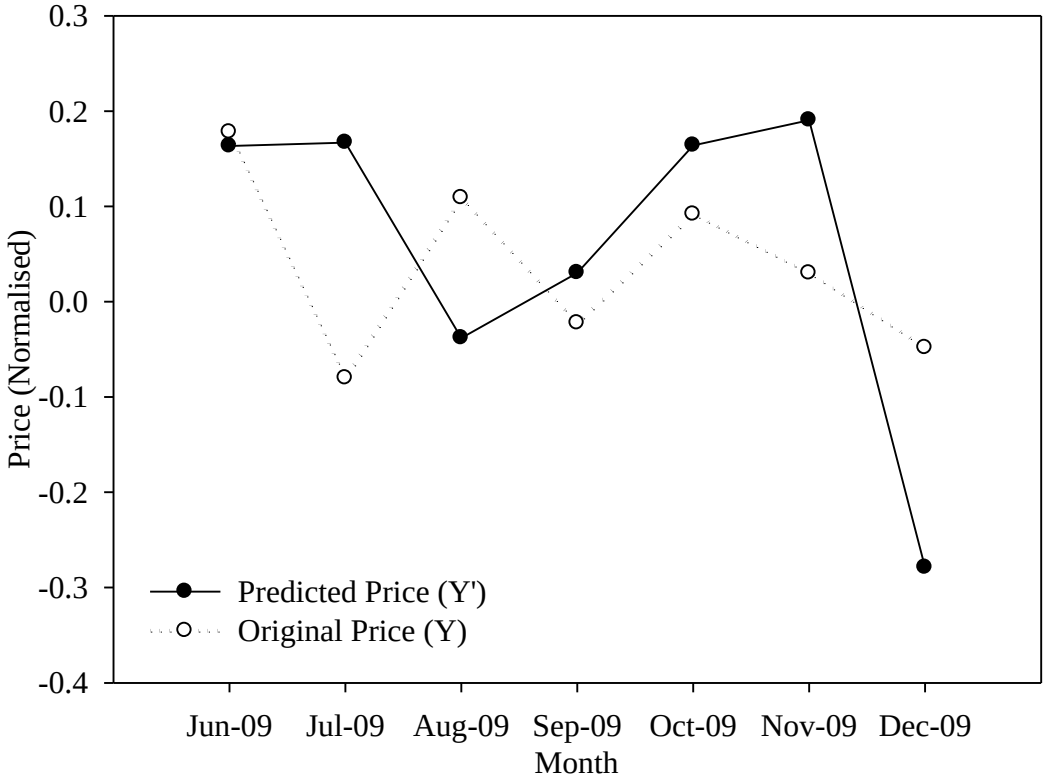


Figure 5.11 The Best Directional Performance Result (85.71%) with 0.969 NMSE for Training Set C, Subset C-1 of the Linguistic-Quantitative (LQ) Prediction Model based on the Directional Statistic (D_{stat}) Value, Trained with a 70:30 Per cent Data Ratio and 4 Hidden Neurons.

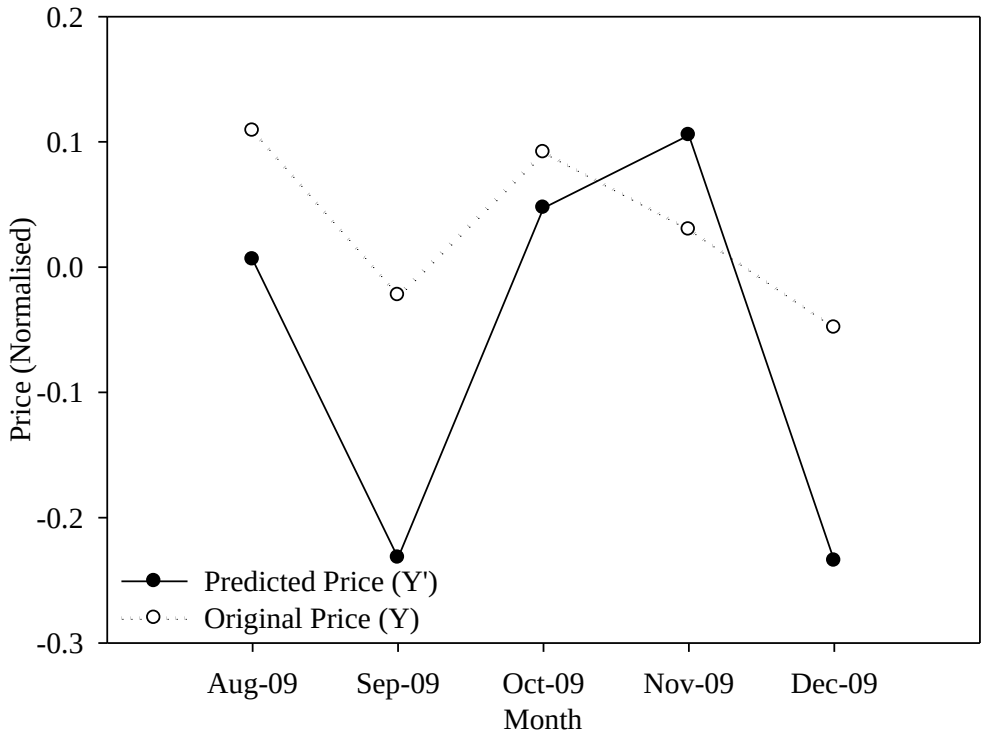


Figure 5.12 The Best Directional Performance Result (80.00%) with 0.938 NMSE for Training Set C, Subset C-2 of the Linguistic-Quantitative (LQ) Prediction Model based on the Directional Statistic (D_{stat}) Value, Trained with an 80:20 Per cent Data Ratio and 5 Hidden Neurons.

Figures 5.13 and 5.14 show the error generalisation based on the maximum D_{stat} obtained from training sets C-1 and C-2. The error was mapped based on the variance calculated between the original prices against the predicted prices, as indicated by the values of the NMSE and the RMSE. In Figures 5.13 and 5.14, the graphs interpret the error span that was determined between the original prices and the predicted prices as true error and false error. An error is considered to be true when the directional gradient of a predicted price is positive and parallel with the gradient of the original price, while a false error is the opposite.

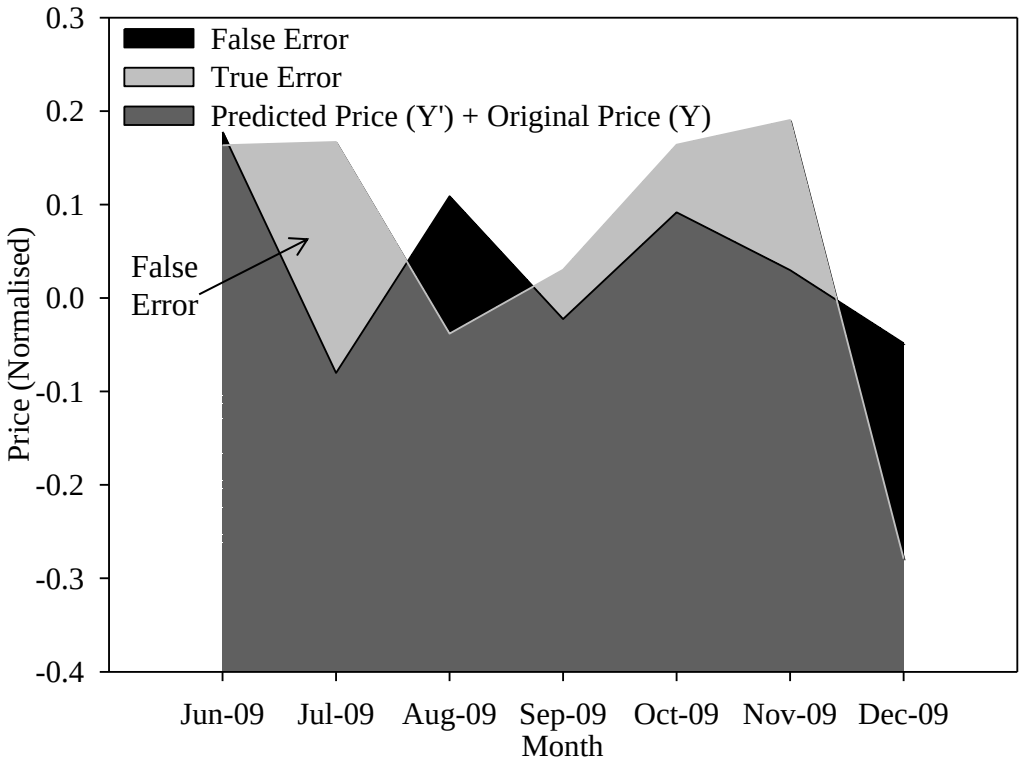


Figure 5.13 The Error Mapping for the Best Directional Performance Result (85.71%) with 0.969 NMSE for Training Set C, Subset C-1 of the Linguistic-Quantitative (LQ) Prediction Model based on Directional Statistic (D_{stat}) Value, Trained with a 70:30 Percent Data Ratio and 4 Hidden Neurons.

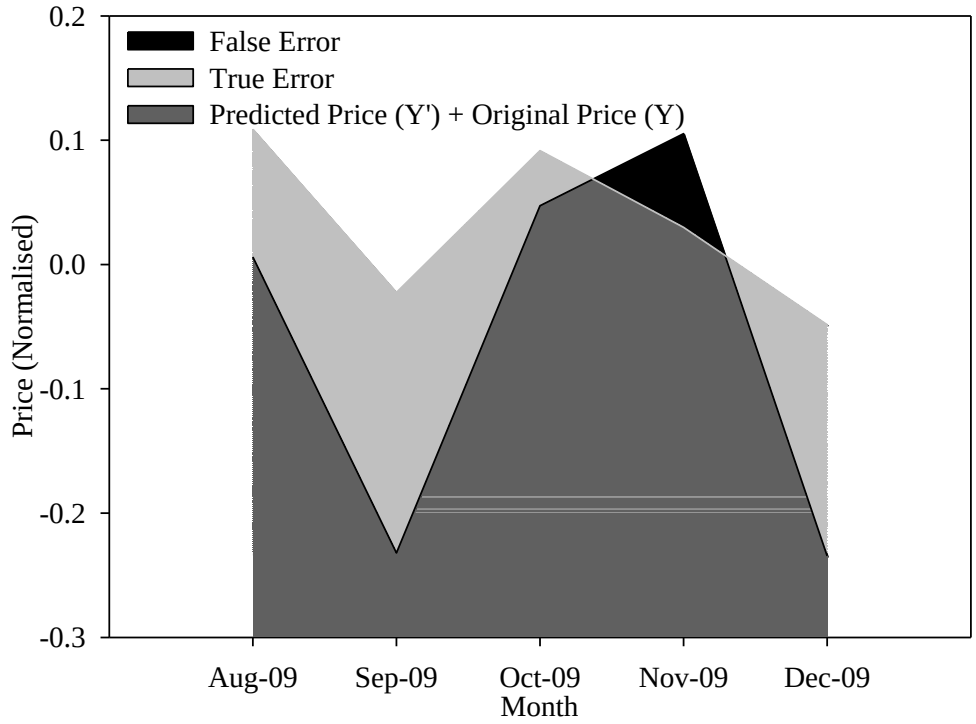


Figure 5.14 The Error Mapping for the Best Directional Performance Result (80.00%) for Training Set C, Subset C-2 of the Linguistic-Quantitative (LQ) Prediction Model based on the Directional Statistic (D_{stat}) Value, Trained with an 80:20 Percent Data Ratio and 5 Hidden Neurons.

An observation was made in Figures 5.11 to 5.14 that analysed the sensitivity of the network in predicting the price with different percentages of data ratios at different durations of the time period than set A and set B. The analysis showed that a bigger percentage of the testing data helped to improve the directional performance of a training subset, as compared to the smaller percentage used for testing data. Nonetheless, a greater percentage of training data helped the network to produce minimal errors in the prediction because the large amount of learning that it processed helped it to derive better reasoning in deriving prediction outcomes.

Subsection 5.6.1.4 will discuss the overall performance of this Linguistic-Quantitative (LQ) prediction model by comparing the training depicted in sets A, B and C of the subsections 5.6.1.1, 5.6.1.2 and 5.6.1.3.

5.6.1.4 Overall Results for the Linguistic-Quantitative (LQ) Prediction Model

The experimental observation produced a spread set of results. The best NMSE values observed in the simulations conducted in section 5.6 with subsections 5.6.1.1 to 5.6.1.3 delivered fairly good outcomes, with 2/3 training sets delivering NMSE values lower than 1.0 marks. The D_{stat} values that were observed as the secondary performance indicator for the linguistic-quantitative (LQ) prediction model also delivered promising results, with training sets A and B achieving averages of 80.00% to 87.10% for the directional accuracy. These performances are indicated in Table 5.26.

Table 5.26 The Best Prediction Performance for the Linguistic-Quantitative (LQ) Prediction Model based on Training Sets A, B, and C.

TRAINING : TESTING RATIO	SET A (JAN '84-DEC '09)			SET B (JAN '06-DEC '09)			SET C (JAN '08-DEC '09)		
	NMSE	RMSE	D*stat* (%)	NMSE	RMSE	D*stat* (%)	NMSE	RMSE	D*stat* (%)
90:10	0.054	0.021	87.10	0.950	0.177	60.00	0.652	0.164	50.00
80:20	0.212	0.075	77.42	1.879	0.156	90.00	0.934	0.162	80.00
70:30	1.107	0.101	77.66	1.295	0.121	85.71	0.969	0.156	85.71

By comparing all the results obtained, the optimal result with the minimum NMSE value and the maximum D_{stat} value were derived from training set A. This set produced 0.054 for its NMSE and 87.10% for its $D_{stat,}$ with a competitive 0.021 RMSE. This validates the hypothesis that linguistic information adds valuable information to and influences the prediction network as discussed in section 5.1 of this chapter. The prediction results with the minimum NMSE value of 0.950 and the maximum D_{stat} value of 90.00% obtained in training set B proved that the data enable the network to learn the function-mapping task, even with relatively limited linguistic data. The results in training set C also proved the network's ability to learn the function-mapping task with a small set of quantitative and linguistic data.

The best performances of each set are presented in Table 5.27, where the best prediction outcomes derived from training subset A-2 are followed by training subsets B-2 and C-1. Next, the best results gained from this Linguistic-Quantitative model, which was trained with the Back-Propagation Neural Networks (BPNN), were compared with other machine-learning approaches discussed in Chapter 2 in order to evaluate the BPNN's performance compared with other techniques.

Table 5.27 The Best Prediction Performance of Training Sets A, B, and C for The Linguistic-Quantitative (LQ) Prediction Model.

SET	TRAINING: TESTING RATIO	HIDDEN NEURONS	ACTIVATION FUNCTION	PERFORMANCE INDICATOR		
				NMSE	RMSE	D*stat* (%)
A-2	90:10	5	Zero-based Log Sigmoid	0.055	0.021	87.10
B-2	70:30	5	Hyperbolic Tangent	0.426	0.181	71.43
C-1	70:30	4	Hyperbolic Tangent	0.969	0.156	85.71

5.6.1.5 Comparison of the Results with Other Machine-Learning Approaches

In order to validate the artificial neural networks (ANN)'s credibility as an accurate time-series prediction tool through its back propagation neural network (BPNN) tool, the performance results discussed in subsections 5.6.1.1 to 5.6.1.4 were compared with other machine-learning approaches in this thesis. The approaches used for this comparison were the support vector machine (SVM), linear regression (LR) and the Gaussian processes (GP), which were discussed in subsection 5.4.2.5. For this comparison, the root mean squared error (RMSE) was used as the principal performance indicator in this section.

Through the evaluation of the results that were based on the same sets of data discussed in subsection 5.5.1 and summarised in Table 5.15, it can be seen that the SVM, LR and GP performed competitively with BPNN. The minimal value of the RMSE achieved by BPNN in training set A-2 puts BPNN in the lead. This verifies the BPNN's credibility as a good prediction tool for large amounts of data. The performance indicators for these experiments are based on the values of the RMSE and the mean absolute error (MAE) obtained from each training set. The MAE is computed by the formula provided in equation 5.10 of this chapter. The performance results of the SVM, LR and GP approaches are laid out in Table 5.28 and presented in Figure 5.15.

Table 5.28 The Results of the Linguistic-Quantitative (LQ) Prediction Model from The Support Vector Machine (SVM), Linear Regressions (LR) and the Gaussian Process (GP) Approaches.

TRAINING SET	BPNN	SVM		LR		GP	
	RMSE	RMSE	MAE	RMSE	MAE	RMSE	MAE
A-2	0.021	0.1168	0.1168	0.0882	0.0700	0.1258	0.1026
B-2	0.181	0.0975	0.1217	0.0680	0.0567	0.0849	0.0686
C-1	0.156	0.0274	0.0254	0.0407	0.0391	0.0396	0.0376

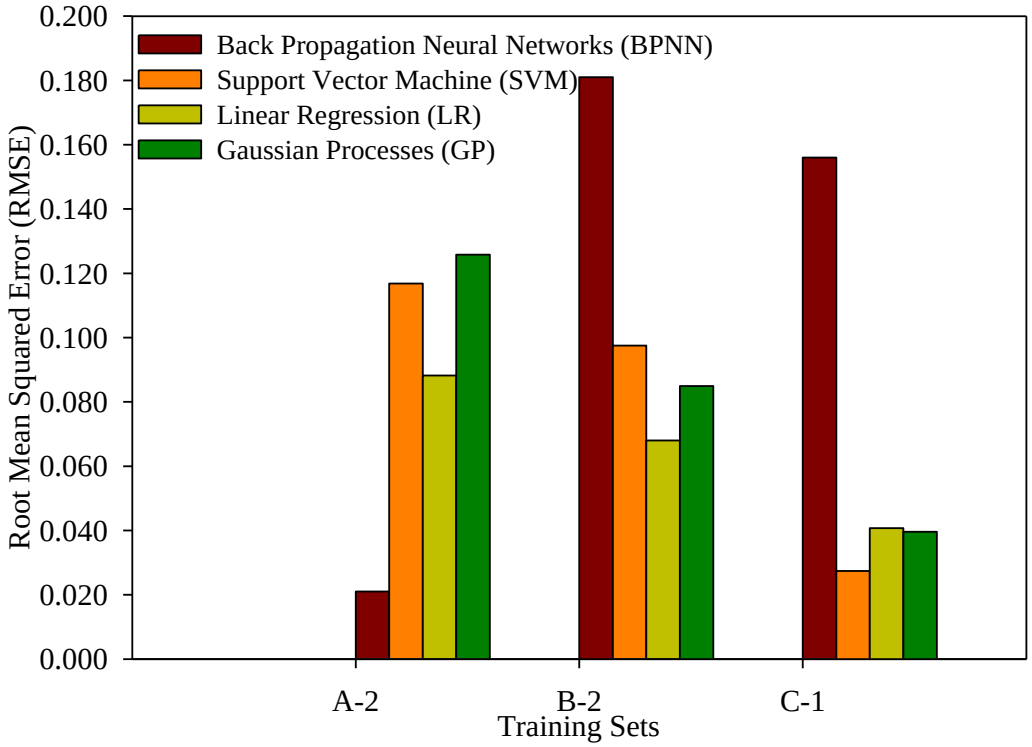


Figure 5.15 The Back Propagation Neural Networks (BPNN) Performance for the Linguistic-Quantitative (LQ) Prediction Model: A Comparison with Other Machine-Learning Approaches Based on Training Sets A-2, B-2 and C-1.

In training sets B-2 and C-1, the BPNN was moderately accurate when compared to the SVM, LR and GP approaches. In reference to Table 5.28, the best prediction performance for training set B-2 was derived from the LR with a minimum RMSE value of 0.0680, followed by the GP and the SVM. Meanwhile, for training set C-1, the minimum RMSE value was obtained from the SVM and was followed by the GP and the LR. The results produced in Table 5.28 prove that the size of the data used in the trainings had an impact on the predictive ability of the BPNN. The BPNN needs sufficient historical information for it to learn and to derive reasoning and accurate prediction. This was shown by the BPNN's good performance results obtained from utilising data in training set A. The BPNN's optimal results generated from training sets A-2, B-2 and C-1 were used as a benchmark for this section. The minimum RMSE values were then chosen from the three training sets mentioned, in order to compare them with the minimum RMSE value from each machine-learning approach

discussed in this section. The optimal results obtained from each approach were then selected and compared with the BPNN in Table 5.29, as visualised in Figure 5.16.

Table 5.29 The Comparison of the Best Results from the Support Vector Machine (SVM), the Linear Regressions (LR) and the Gaussian Process (GP) Approaches with the Back-Propagation Neural Networks (BPNN) used in the Linguistic-Quantitative (LQ) Prediction Model.

MACHINE LEARNING APPROACH	**PERFORMANCE INDICATOR RMSE**
BPNN	0.021
SVM	0.027
LR	0.041
GP	0.040

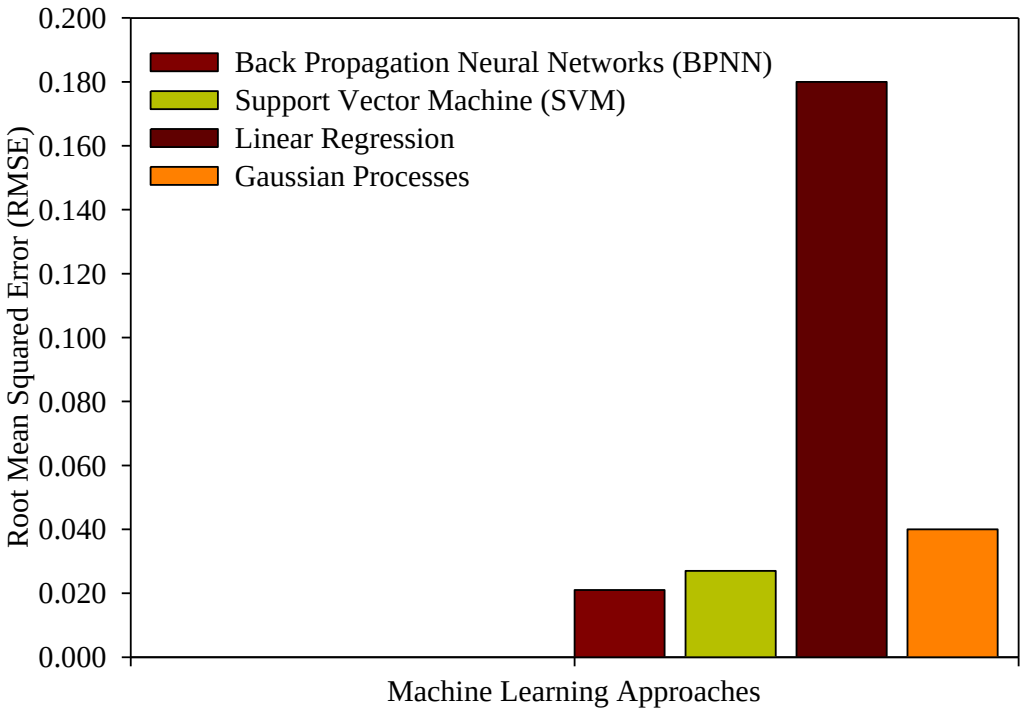


Figure 5.16 The Back- Propagation Neural Networks (BPNN) Performance for the Linguistic-Quantitative (LQ) Prediction Model Based on the Minimum RMSE Value: A Comparison with Other Machine-Learning Approaches.

Based on the evaluation made in Table 5.29 and the performance comparison in Table 5.16, the BPNN leads the other machine-learning approaches in terms of its competitive RMSE value, followed by the competitive SVM and GP, with the LR in the last position. This result proves that the integration of linguistic information with a quantitative prediction model adds value to the model. The summary of the discussions and the simulation results in this chapter is discussed in section 5.7.

5.7 CONCLUSION

In this chapter, the elements of the main theories were discussed, as follows:

i) the exploitation of news sentiments as inputs for predicting the crude oil price;
ii) the development of an ANN model predicting the crude oil price from linguistic information;
iii) the hybridisation of both linguistic and quantitative data to predict the crude oil price; and
iv) the empirical results derived from the linguistic prediction model and the linguistic-quantitative prediction model.

From the discussion regarding the exploitation of news sentiments as input, we concluded that the linguistic element adds information to the crude oil price prediction models. The exploitation of this kind of information gives insight into the reaction of the market and the impact on the price. The qualitative information extracted from the mined news articles complemented the quantitative element in the network, as it indicated events that happened quantitatively. The linguistic information is also capable of adding information, even with incomplete or small sets of data. This has been proven by the outcomes obtained from the simulation tests. The results obtained and discussed in subsection 5.6.1.5 proved the added credibility of BPNN as a reliable prediction tool in addition to other machine-learning approaches. In order to improve the BPNN's reliability in providing promising prediction outcomes, the utilisation of an extensive amount of data is essential. Nonetheless, the simulations made in subsection 5.6.1.5 on small set of data also proved to produce accurate results. The optimal results obtained from both the linguistic prediction model and the LQ model are summarised in Table 5.30 as follows:

Table 5.30 The Summary of the Optimal Performance for Models in Chapter 5.

MODEL	**NMSE**	**RMSE**	**D*stat* (%)**
Linguistic prediction model:			
Directional price	-	0.081	80.00
Discrete price (normalised)	0.685	-	100.00
Linguistic-quantitative (LQ) prediction model	0.055	0.021	87.10

Chapter 6 Conclusion and Suggestions for Future Research

Overview

This chapter summarises the main achievement of this research. It also identifies the main contributions of the mathematical models developed in this thesis. Finally, it suggests some improvements that could be made to the existing models, and the future potential research topics that might be of interest for further investigation in the future.

6.1 INTRODUCTION

The crude oil price market is known for its obscurity and complexity. Due to this high uncertainty aligned to the irregular events in the market, predicting the market's behaviour is a challenging task. Nevertheless, its uncertainty and volatile characteristics attract much attention of researchers.

A hybrid of machine-learning and computational intelligence approach are combined to utilise historical quantitative data with linguistic elements from online news services was proposed to predict prices. The models developed in this research were presented as five different mathematical models: the hierarchical conceptual (HC) model; the ANN-Q quantitative model; the rule-based expert model; the linguistic model; and lastly, the hybridisation of linguistic and quantitative (LQ) model.

The hierarchical conceptual model discussed in Chapter 3 was built as a basis of information retrieval to understand the crude oil price market's behaviour. Through the development of this model, a systematic approach was proposed to determine the main contributors to this volatile behaviour. The main contributors to this volatile behaviour as extracted by the HC model in chapter 3 are tabulated in Table 6.1.

Table 6.1 The Key Impact Factors of Crude Oil Market.

VARIABLES	FACTORS
$\boldsymbol{S^{186T}}$	**SUPPLY**
S_{a1}	Productions of OPEC countries
S_{a2}	Productions of Non-OPEC countries
S_{b1}	Proved reserves of OPEC countries
S_{b2}	Proved reserves of OECD countries
S_{b3}	Number of well drilled
$\boldsymbol{D^{T}}$	**DEMAND**
D_{a1}	Consumption of OECD countries
D_{a2}	Consumption of China
D_{a3}	Consumption of India
$\boldsymbol{I^{T}}$	**INVENTORY**
I_{a1}	Ending stocks of OECD countries
I_{a2}	Ending stocks of US
I_{b1}	US petroleum imports from OPEC countries
I_{b2}	US petroleum imports from Non-OPEC countries
I_{c1}	US crude oil imports from OPEC countries
I_{c2}	US crude oil imports from Non-OPEC countries
$\boldsymbol{E^{T}}$	**ECONOMY**
E_{a1}	Foreign Exchange of GBP/USD
E_{a2}	Foreign Exchange of Yen/USD
E_{a3}	Foreign Exchange of Euro/USD
E_{b1}	US Growth Domestic Products (GDP)
E_{c1}	US Inflation rate
E_{d1}	US Consumer Price Index (CPI)
$\boldsymbol{P^{T}}$	**POPULATIONS**
P_{a1}	Population of developed countries
P_{a2}	Population of less developed countries
WTI	**WEST TEXAS INTERMEDIATE PRICE**

[186] T is the accumulated value or the core contributor of each feature.

A price classification was then proposed to obtain this information based on the turning points of the price in the market. The temporal price with large turning points is believed to contain sufficient information that contributes to the volatile behaviour of the market. A benchmark price for crude oil, the West Texas Intermediate (WTI), was utilised in the HC model as a reference for this purpose. The price data obtained from credible sources discussed in subsection 3.3.1 were classed into High and Low cases to denote the upward and downward price movement in which information was revealed by the variance between the previous and current data. Data from these large turning points based on the variances were then stored and used as a basis to obtain the knowledge of market's behaviour through a search of related articles from the Google News service.

The online news articles were mined using the keywords "crude oil", "oil price", and "crude oil price" from the service, which correspond to the features extraction objective discussed in subsection 3.2.3. Initially, the online news articles were mined to examine the rules contained in the articles, which reflect the market's volatile behaviour. The rules contained in the articles were extracted, with the main features (the impact factors) of the rules selected, based on their frequency of occurrences in the articles. These extracted impact factors were collected, documented and classed in accordance to their level of contribution, which are aligned with a hierarchy of importance. The HC model had successfully derived a number of impact factors that were initially hypothesised to be the main contributors to the market fluctuation. The impact factors were used as input in the quantitative prediction model to prove the initial hypothesis.

Also discussed in Chapter 3 is the development of artificial neural network-quantitative prediction model (ANN-Q) which, among other objectives in the chapter, validates the factors retrieved from the HC model. The validation was made by using factors retrieved from the HC model as input, and by conducting a sensitivity analysis with different combinations of input used in the network. The input combination training with the minimum error demonstrates dependency of the price on the input variables as suggested by [26]. This model was proposed to introduce a different method of predicting the crude oil price, different than

conventional statistics and econometric techniques by providing a more accurate prediction results. To materialise the more precise prediction outcomes, a collection of significant inputs for training is crucial. Hence, the HC model acted as a basis for the process. Data collection based on input obtained from the HC model was used in this model to train the network and to validate its generalisation ability.

An artificial neural network (ANN) tool trained with back propagation neural network (BPNN) was employed to the model and trained with 3, 4 and 5 hidden neurons and with 90:10 percent, 80:20 percent and 70:30 percent of training: testing data ratio. The training data enabled the network to learn the patterns present in the input data and deliver the best prediction forecast. Data convergence was established from the inputs provision delivered by the HC model, evidenced by the preliminary results obtained in the training. The model offered promising prediction results, which were derived from the employment of all inputs provided by HC model, validating the credibility of the HC model and the ANN as the prediction tool. The application demonstrates accurate prediction outcomes, evidenced by the minimum value of error (NMSE: 0.00896, RMSE: 2.26900) and optimum value of directional accuracy (D_{stat}*:* 93.33%) as discussed in subsection 3.4.1.

The content utilisation and sentiment analysis applied with the fuzzy grammar extraction module in Chapter 4 is a core contributor to the research. Chapter 4 is an extension of the contribution demonstrated by the HC model in Chapter 3. Chapter 4 demonstrates a more systematic approach of information retrieval. It employs content utilisation and sentiment mining from news articles with the application of fuzzy grammar fragment extraction to retrieve the sentiments of the market. In Chapter 4, a 'dictionary' that contains grammar definitions was built and customised in accordance with the crude oil market. The dictionary that represents the terminal grammar was established with the objective of deriving grammar rules. The importance of the selection of the appropriate definitions has helped to derive relevant grammar fragments from news articles for sentiment mining and analysis. A grammar combination of one or more impact factors (*coreTerm = oilFactor)* aligned with one or more *contentCategory* (*positiveTerm, negativeTerm)* in a sentence of an article is the rule for grammar fragment extraction. The *oilFactor* and

contentCategory that denote the noun and verb of an English language structure were used to mine sentiments from the news article that reflect price movement and its dependent input. These sentiments are not only useful for analysis, but are also used as the main inputs to construct a decision tree, the linguistic input for the fuzzy expert system, and the sole input used in the linguistic prediction model. Rules were extracted eliciting the 25 sets of rules that explain the market's price behaviour, acting as the 'expert' rule base, storing and mapping the behaviour of the market.

A linguistic prediction model which utilised an ANN whose aim was to integrate linguistic information into a quantitative prediction model was established in Chapter 5. This model was built to validate that the extracted inputs obtained from grammar fragments from the news sentiment mining and analysis did affect the price market. Monthly quantified and aggregated sentiments were used to train this model to recognise the temporal pattern in market data. In the linguistic prediction model, linguistic elements that were obtained from Chapter 4 were used to predict the price in three ways: directional price, normalised price and original price.

These quantified sentiments were later employed as a complementary input to the quantitative information in the linguistic-quantitative (LQ) model in section 5.5 of Chapter 5. The results obtained from the LQ model demonstrate that linguistic information adds value to the prediction network, even with small data. This also proved the hypothesis that linguistic information improved the performance of a model, as discussed through other applications in Chapter 5's literature. This model also adds credibility to ANN as a promising prediction tool as it is capable of function mapping the input used in the network even with missing data, as evidenced by its minimum error value (NMSE: 0.055, RMSE: 0.021) and the optimum directional accuracy degree (D_{stat}: 87.10%).

6.2 MAIN CONTRIBUTIONS OF THE THESIS

The main contribution of this thesis lies explicitly on the extraction of data used as inputs in each of its model. Appropriate selection of data through a systematic approach employed in HC model and the rule-based expert model were evidenced by the promising prediction outcomes resulting to the inputs interconnected in the network. The interconnections of inputs in the network were determined based on the minimal error value that the prediction models produced. A minimum error value that is closer to zero interprets the high dependency of the price on the input data. The process began with a manual rules extraction process in the HC model to retrieve the information related to the market's behaviour. This approach demonstrates the important contribution to the thesis by contributing the first platform in order to build other models in the thesis. The market features discovered from the HC model has led to the production of new vectors of inputs used by models throughout the research.

This input data 'propagation' contributes to a systematic investigation into the market's behaviour; which begins with the extraction of main features that leads to an understanding of the market through sentiment mining, which is then exploited as the quantifiable linguistic element in anticipating price movements. Nevertheless, the contributions for each model are presented in the following subsections:

6.2.1 Hierarchical Conceptual Model and Artificial Neural Network-Quantitative (ANN-Q) Prediction Model

The contributions of the hierarchical conceptual (HC) model discussed in Chapter 3 are presented as follows:

1. This proposed systematic approach has helped to identify the oil market's main impact factors that contribute to the volatility of the crude oil price. This was evaluated based on the prediction accuracy that it produced from the usage of these data in the quantitative prediction model. The effectiveness and appropriateness of this data selection has helped to deliberate extensive amount of inputs for the quantitative model which later validated as best training input data for the model;
2. The features found from the mined news articles through hierarchical conceptual information extraction, proved to present a fair observation to the market behaviour and its domino effects on the volatilities of the price;
3. All inputs discovered through data classification proposed by this model, proved to generate good associations of data that was evidenced by the promising prediction result based on the minimum error and the directional accuracy obtained from the quantitative model; and
4. The promising prediction results obtained from this model confirmed the data classification process proposed, that leads to the selection of the appropriate data, then to be used as inputs in the model.

6.2.2 Content Utilisation and Sentiment Analysis with Fuzzy Grammar Fragment Extraction Application for Rule-based Expert Model

The contributions of the content utilisation and sentiment analysis with fuzzy grammar fragment extraction application for rule-based expert model presented in Chapter 4 are as follows:

1. The process of constructing a grammar definition through grammar building, based on language structure with Context-Free Grammar (CFG) was developed. This domain-specific process has led to the construction of grammar that was then used to annotate and extract the grammar combinations of a *coreTerm* and a *contentCategory* from a fragment of an article, utilising sentiment mining and analysis.
2. Extracted grammar fragments derived in point (1) had managed to extract important input data for a rule-based expert model and the linguistic prediction model in Chapter 5. The monthly aggregated sentiments derived from the extracted grammar fragments were then implemented as the sole input into the linguistic prediction model. The quantified sentiments, obtained from the grammar fragments, were used as input to delineate the rules in a decision tree discussed in section 4.7.2.
3. The tree produced in (2) (the second contribution) was exploited in a fuzzy inference system as input to form the rules systematically and to establish the expert system in section 4.8.

6.2.3 Linguistic Prediction Model and Linguistic-Quantitative (LQ) Hybrid Model

The contributions of the linguistic prediction model and linguistic-quantitative (LQ) prediction model discussed in Chapter 5 are presented as follows:

1. News sentiments obtained from Chapter 4, exploited as the sole linguistic input, encoded impacts of the market by producing promising results from the linguistic prediction model. The prediction results demonstrated that linguistic information is a good element and a dependent factor on the crude oil price.
2. The linguistic information contained in the news sentiments complements the quantitative values contained in the quantitative model discussed in section 5.5 of Chapter 5, where it capable of adding valuable information into the hybrid (LQ) prediction model even with small size data. A series of incomplete (missing) and small size linguistic data were fed into the network to train with the quantitative data for generalisation. This confirms the credibility of the domain-specific grammar definitions in Chapter 4 as a valid extractor of information.
3. The integration of both quantitative and linguistic data into a hybrid prediction model imitates the routine of a crude oil market's regulator in calculating risk of actions, which is based on the 'facts' (quantitative data) and 'rumours' (linguistic data) received in the market. The model imitates the regulator's anticipations before executing a price hedge in the market. The integration of both types of data provides a comprehensive learning set (experience) to the neural network in order to propagate and learn, and then establishes an 'informed' decision to hedge price from the prediction.

6.3 SUGGESTIONS FOR FURTHER RESEARCH

The suggestions for further research are divided into suggestions that focus on some possible refinements to the models proposed in thesis, and suggestions of any new directions for research related to crude oil price prediction.

Some improvements that could be made to the research are presented in (1) with suggestion for potential future research in (2)-(4):

1. **An automatic approach** to gather fragments that are extracted from articles and placed into a systematic database would be an appealing improvement. The extraction process of these grammar fragments from articles is laborious since it involved analysing thousands of grammar fragments. The implementation of an automated approach will save more time and will channel more energy towards the exploration of different and other aspects of the research.

2. **A short-term prediction model (technique).** After a long time-term prediction model is implemented, an accurate short-term period predictive model would be the next research area to explore.

3. **An evolving and online short-term prediction model.** By exploiting the knowledge gathered in the research, a useful online short-term price forecasting model could be developed to provide a daily or weekly price. This would be an attractive area to explore as it will provide current and fast analysis for practitioners to react in accordance to price market volatility. Furthermore, vast fluctuations in daily prices may enable new factors to be extracted, providing an interesting area to explore.

3. **A fuzzy system learning approach.** A news mining approach to mining the fuzzy structure online as an extension to the fuzzy expert system as discussed in section 4.8 would be an interesting area to explore. A system that is not just helping to predict the market, but also helps to make decisions would be an appealing area to study. As investors use their anticipation skill based on ‘rumours’ to hedge the price in the market, a neuro-fuzzy approach with an adaptive neuro-fuzzy inference system (ANFIS) is a good model to explore and propose. This can establish a decision-

making system that extracts information (input) online and compute it in the system automatically, thus producing a decision (output) based on information from the extracted crude oil market's rules.

References

[1] R. Takeishi, "Economic Topics: Background of the Increase in Oil Prices and Prospects for the Future," Fujitsu Research Institute, 27 January 2007. [Online]. Available: http://jp.fujitsu.com/group/fri/en/column/economic-topics/2007/2007-01-11-3.html. [Accessed 20 January 2013].

[2] "Short-term Energy Outlook- STEO Supplement: Why are oil prices so high?," International Energy Agency (IEA), [Online]. Available: http://www.eia.gov/forecasts/steo/special/pdf/high-oil-price.pdf. [Accessed 6 July 2013].

[3] "NYMEX Daily Reports," CME Group Inc., [Online]. Available: http://www.cmegroup.com/trading/energy/nymex-daily-reports.html. [Accessed 15 October 2008].

[4] A. Shah, "Top 10 Most Traded Commodities in the World," *Digital Stockmarket: Global Stock Watch,* 22 November 2011.

[5] "Caltex: Determining Fuel Prices," Chevron Corporation, [Online]. Available: http://www.caltex.com/global/resources/determining-fuel-prices/. [Accessed 16 January 2009].

[6] "FREQUENTLY ASKED QUESTIONS: What are the differences between various types of crude oil prices?," US Energy Information Administration (EIA), [Online]. Available: http://www.eia.gov/tools/faqs/faq.cfm?id=11&t=5. [Accessed 10 August 2013].

[7] L. Sheridan, "Aljazeera: Oil prices to rise again- As Opec Meets in Vienna, an Industry Expert Says the Recent Slump May Not Last," 10 September 2008. [Online]. Available: http://www.aljazeera.com/focus/2008/09/200898133143509358.html. [Accessed 25 May 2009].

[8] “Economagic.com: Economic Time Series Page,” [Online]. Available: http://www.economagic.com. [Accessed 1 December 2008].

[9] “WRTG Economics,” [Online]. Available: http://wtrg.com/prices.htm. [Accessed 28 August 2009].

[10] “US Energy Information Administration (EIA): Independent Statistics and Analysis,” [Online]. Available: http://www.eia.gov/. [Accessed 25 November 2011].

[11] “MLA: "Futures Market Basics - CFTC - U.S. Commodity Futures Trading",” [Online]. Available: http://www.cftc.gov/ConsumerProtection/EducationCenter/FuturesMarketBasics/index.htm. [Accessed 17 September 2013].

[12] K. Amadeo, “US Economy: News & Issues,” *How are Oil Prices Determined?,* 22 February 2012.

[13] Y. Nelson, S. Stoner and G. Gemis, “Results of Delphi VIII Survey of Oil Price Forecasts,” in *Energy Report*, California, California Energy Commission, 1994.

[14] “The World Bank: Working for a World Free of Poverty,” [Online]. Available: http://data.worldbank.org/. [Accessed 4 January 2009].

[15] International Energy Agency, “Key World Energy Statistics,” 2012. [Online]. Available: http://www.iea.org/publications/freepublications/publication/kwes.pdf. [Accessed 5 July 2013].

[16] “The World Factbook,” Central of Intelligence Agency (CIA), [Online]. Available: https://www.cia.gov/library/publications/the-world-factbook/rankorder/2241rank.html. [Accessed 13 September 2013].

[17] E. Bummiler and T. Shanker, “U.S. Sends Top Iranian Leader a Warning on Strait Threat,” New York Times, 12 Jnauary 2012. [Online]. Available:

http://www.nytimes.com/2012/01/13/world/middleeast/us-warns-top-iran-leader-not-to-shut-strait-of-hormuz.html. [Accessed 15 March 2013].

[18] "Special Report: Hurricane Katrina's Impact on U.S. Energy," US Energy Information Administration (EIA), 31 August 2005. [Online]. Available: http://www.eia.gov/oog/special/eia1_katrina_083105.html. [Accessed 16 June 2013].

[19] S. P. A. Brown, R. Virmani and R. Alm, "Crude Awakening: Behind the Surge in Oil Prices," *Economic Letter-Insights from the Federal Reserve Bank of Dallas,* vol. 3, no. 5, 2008.

[20] M. Ye, J. Zyren, J. Shore, "A Monthly Crude Oil Price Forecasting Model using Relative Inventories," *International Journal of Forecasting,* vol. 21, pp. 491-501, 2006.

[21] F. Birol, "Analysis of the Impact of High Oil Price on the Global Economy," *International Energy Agency,* 2004.

[22] S. Ghosh, "Import Demand of Crude Oil and Economic Growth: Evidence from India," *Energy Policy,* vol. 37, pp. 699-702, 2009.

[23] H. Shin, T. Hou, K. Park, CK. Park, S. Choi, "Prediction of movement direction in crude oil prices based on semi-supervised learning," *Decision Support Systems,* vol. 50, pp. 164-175, 2012.

[24] K. He, K. K. Lai, S-M. Guu, J. Zhang, "A wavelet based multi scale var model for agricultural market," in *Modelling, Comptation and Optimisation in Information Systems and Management Sciences*, Berlin, Germany, 2008.

[25] "Daily Forex Signals & Analysis," [Online]. Available: http://www.meta4forexbroker.com/2011/09/what-is-brent-crude/. [Accessed 20 Jun 2013].

[26] M. Negnevitsky, Artificial Intelligence: A Guide to Intelligent System, Essex:

Pearson Education, 2005.

[27] E. Panas, V. Ninni, “Are Oil Markets Chaotic? A Nonlinear Dynamic Analysis,” *Energy Economics,* vol. 22, pp. 549-568, 2000.

[28] T. Bollerslev, “Generalized Autoregressive Conditional Heteroskedasticity,” *Journal of Econometrics,* vol. 31, pp. 307-327, 1986.

[29] Z. Bar-Yossef, A. Berg, et. al, “Approximating Aggregate Queries about Web Pages via Random Walks,” in *Proceedings of the 26th VLDB Conference*, Egypt, 2000.

[30] “Random walk model,” Decision 411 Company, [Online]. Available: http://people.duke.edu/~rnau/411rand.htm. [Accessed 20 May 2013].

[31] B. Abramson, A. Finniza, “Using Belief Networks to Forecast Oil Prices,” *International Journal of Forecasting,* vol. 7, pp. 299-315, 1991.

[32] B. Abramson, A. Finizza, “Probabilistic Forecasts from Probabilistic Models: A Case Study in the Oil Market,” *International Journal of Forecasting,* vol. 11, pp. 63-72, 1995.

[33] C. Morana, “A Semiparametric Approach to a Short Term Oil Price Forecasting,” *Energy Economics,* vol. 23, pp. 325-338, 2001.

[34] G. Barone-Adesi, F. Bourgoin, K. Giannopoulus, “Don't Look Back,” *Risk,* pp. 100-104, 11 August 1998.

[35] “A Primer on Martingales,” Petrov Financial, [Online]. Available: http://www.petrovfinancial.com/?page_id=880. [Accessed 11 November 2012].

[36] W. Xie, L. Yu, S. Xu, S. Wang, “A New Method for Crude Oil Price Forecasting Based on Support Vector Machines,” in *ICCS'06 Proceedings of the 6th International Conference on Computational Science*, Berlin, Heidelberg, 2006.

[37] S.N. Abdullah, X.Zeng, “Machine Learning Approach for Crude Oil Price Prediction with Artificial Neural Network-Quantitative (ANN-Q) Model,” in *IEEE International Joint Conference for Neural Networks*, Barcelona, 2010.

[38] J.W.S. Hu, Y.C. Hu, R. Lin, “Applying Neural Networks to Prices Prediction of Crude Oil Futures,” *Mathematical Problems in Engineering*, vol. 2012, pp. 1-12, 2012.

[39] S. Wang, L. Yu, K.K. Lai, “Crude Oil Price Forecasting with TEI@I Methodology,” *Journal of Systems Science and Complexity*, vol. 18, no. 2, pp. 145-165, 2005.

[40] I. Ginzburg, U. Naftaly, D. Horn, N. Intrator, “Averaged and Decorrelated Neural Networks as a Time Series Predictor,” in *International Conference for Pattern Recognition, 12th ICPR*, 1994.

[41] L. Yu, S. Wang, K. K. Lai, J. Yen, “Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm,” *Energy Economics*, vol. 30, no. 5, pp. 2623-2635, 2008.

[42] L. Yu, S. Wang, K.K. Lai, “A Rough-Set-Refined Text Mining Approach for Crude Oil Market Tendency Forecasting,” *International Journal of Knowledge System Sciences*, vol. 2, no. 1, 2005.

[43] S. Wang, L. Yu, K. K. Lai, “An EMD-Based Neural Network Ensemble Learning Model for World Crude Oil Spot Price Forecasting,” *Soft Computing Applications in Business*, vol. 230, pp. 261-271, 2008.

[44] W. Xie, L. Yu, S. Xu, “A New Method for Crude Oil Price Forecasting based on Support Vector Machine,” in *Springer*, Heidelberg, 2006.

[45] J. Leigh, “Economic Meltdown in America Saves the World from Peak Oil,” *Energy Bulletin*, 9 October 2009.

[46] D. Pylypczak, “Top 5 Global Oil Stocks by Market Cap,” *Commodity HQ*, 7

October 2012.

[47] Z. Tang, P.A. Fishwick, "Feedforward Neural Networks as Models for Time Series Forecasting," *ORSA Journal on Computing,* vol. 5, no. 4, pp. 374-385, 1993.

[48] V. Rao, H. Rao, "C++ Neural Networks Application in Financial Asset Management," *Neural Computing and Application Journal*, vol. 2, pp. 13-39, 1995.

[49] "International Energy Agency," [Online]. Available: http://www.iea.org. [Accessed 25 November 2011].

[50] "World Energy Council," [Online]. Available: http://www.worldenergy.org/. [Accessed 2 December 2011].

[51] "Population Reference Bureau," [Online]. Available: http://www.prb.org/. [Accessed 13 January 2012].

[52] "OECD: Better Policies for Better Lives," [Online]. Available: http://www.oecd.org/home/. [Accessed 17 January 2009].

[53] M. F. Nasrudin, "A Model and Application Development for KLSE Stock Market Prediction using Back Propagation Neural Network," in *Msc. Thesis*, Faculty of Technology and Science Information, The National University of Malaysia, 2001.

[54] E. R. Rene, M. E. López, J. H. Kim, and H. S. Park, "Back Propagation Neural Network Model for Predicting the Performance of Immobilized Cell Biofilters Handling Gas-Phase Hydrogen Sulphide and Ammonia," *BioMed Research International,* vol. 2013, no. Article ID 463401, p. 9, 2013.

[55] D.F. Cook & R.E. Shannon, "A Sensitivity Analysis of a Back-propagation Neural Network for Manufacturing Process Parameter," *Journal of Intelligent Manufacturing,* vol. 2, pp. 155-164, 1991.

[56] H. R. Maier and G. C. Dandy, “The Effect of Internal Parameters and Geometry on the Performance of Back-propagation Neural Networks: An Empirical Study,” *Environmental Modelling & Software,* vol. 13, p. 193–209, 1998.

[57] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning Representations by Back-propagating Errors,” *Nature,* vol. 323, no. 6088, pp. 533-536, 1986.

[58] M. A. Haider, K. Pakshirajan, A. Singh, and S. Chaudhry, “Artificial Neural Network-genetic Algorithm Approach to Optimize Media Constituents for Enhancing Lipase Production By A Soil Microorganism,” *Applied Biochemistry and Biotechnology,* vol. 144, no. 3, pp. 225-235, 2008.

[59] P. Kroha, R. Baeza-Yates and B. Kreller, “Text Mining of Business News for Forecasting,” in *17th International Conference on Database and Expert Systems Application (DEXA'06)*, 2006.

[60] V. Milea, R.J. Almeida, U. Kaymak, F. Frasincar, “A Fuzzy Model of the MSCI EURO Index Based on Content Analysis of European Central Bank Statement,” in *2010 IEEE World Congress on Computational Intelligence (WCCI 2010)*, Barcelona, 2010.

[61] “Eurex Exchange,” [Online]. Available: http://www.eurexchange.com/exchange-en/products/idx/msc/432714/. [Accessed 10 November 2013].

[62] R. P. Schumaker and H. Chen, “Textual Analysis of Stock Market Prediction using Breaking Financial News: The AZFinText System,” *ACM Transactions on Information Systems,* vol. 27, no. 2, 2009.

[63] “Google News,” [Online]. Available: http://news.google.co.uk/intl/en_uk/about_google_news.html. [Accessed 15 December 2012].

[64] J. Gombiner, “Carbon Footprinting the Internet,” *Consilience: The Journal of*

Sustainable Development, vol. 5, no. 1, pp. 119-124, 2011.

[65] E. Kuhn, "CNN Politics- Political Ticker: Google Unveil Top Political Searches of 2009," 2009. [Online]. Available: http://politicalticker.blogs.cnn.com/2009/12/18/google-unveils-top-political-searches-of-2009/. [Accessed 15 September 2012].

[66] "The Porter Stemming Algorithm," [Online]. Available: http://tartarus.org/martin/PorterStemmer/. [Accessed 14 September 2012].

[67] "General Inquirer," Harvard University, [Online]. Available: http://www.wjh.harvard.edu/~inquirer/. [Accessed 3 February 2011].

[68] H.D. Klingemann, P.P. Mohler and R.P. Weber, "Das Reichtumsthema in den Thronreden des Kaisers und die Okonomische Entwicklung in Deutschland 1871-1914, Computerunterstutzte Inhaltsanalyse in der empirischen Sozialforschung," Kronberg: Athenaum, 1982.

[69] J. Zhang, Y. Kawai, T. Kumamoto and K. Tanaka, " A Novel Visualization Method for Distinction of Web News Sentiment," *10th International Conference on Web Information Systems Engineering (WISE 2009),*, pp. pp. 181-194, 2009.

[70] D.C. Dunphy, C.G. Bullard and E.E.M. Crossing, " Validation of the General Inquirer Harvard IV Dictionary," in *1974 Pisa Conference on Content Analysis*, 1974.

[71] H.D. Laswell and J.Z. Namenwirth, The Laswell Value Dictionary, New Haven: Yale University Press, 1968.

[72] J.Z. Namenwirth and R.P. Weber, "The Laswell Value Dictionary," in *Pisa Conference on Content Analysis*, Pisa, 1974.

[73] "Wilhelm Maximilian Wundt," [Online]. Available: http://plato.stanford.edu/entries/wilhelm-wundt/. [Accessed 4 November 2012].

[74] N. Chomsky, “Three Models for the Description of Language,” *IRE Transactions on Information Theory*, vol. 2, no. 3, p. 113–124., 1956.

[75] N. Chomsky, “On Certain Formal Properties of Grammars,” *Information and Control*, vol. 2, no. 2, p. 137–167, 1959.

[76] N. Chomsky and M.P Schützenberger, “The Algebraic Theory of Context Free Languages,” in *Computer Programming and Formal Languages*, Amsterdam, 1963.

[77] J.W. Backus, “The Syntax and Semantics of the Proposed International Algebraic Language of the Zurich ACM-GAMM Conference,” in *Proceedings of the International Conference on Information Processing*, UNESCO, 1959.

[78] C.-S. Lee, Y.-F. Kao, Y.-H. Kuo, “Automated ontology construction for unstructured text documents,” *Data Knowledge Engineering*, vol. 60, no. 3, pp. 547-566, 2007.

[79] N. Mohd Sharef and Y. Shen, “Text Fragment Extraction Using Incremental Evolving Fuzzy Grammar Fragments Learner,” in *2010 IEEE World Congress on Computational Intelligence (WCCI 2010)*, Barcelona, 2010.

[80] N. Mohd Sharef, “Text Fragment Identification with Evolving Fuzzy Grammars,” in *PhD Thesis*, Department of Engineering Mathematics, Faculty of Engineering, University of Bristol., 2010.

[81] A. Andreevskaia and S. Bergler, “Mining WordNet for Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses,” in *Proceedings the 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, Stroudsburg, 2006.

[82] A. Esuli and F. Sebastiani, “Determining Term Subjectivity and Term Orientation for Opinion Mining,” in *Proceedings the 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, Stroudsburg, 2006.

[83] H. Sakaji, H. Sakai and S. Masuyama, “Automatic Extraction of Basis Expressions that Indicate Economic Trend,” *Advances in Knowledge Discovery and Data Mining,* pp. 977-984, 2008.

[84] S. A. Ghallab, N. Badr and M. Hashem, “A Fuzzy Expert System For Petroleum Prediction,” in *7th European Computing Conference (ECC '13)*, Dubrovnik, 2013.

[85] X. Zhang, “Crude Oil Price Forecasting using Fuzzy Time Series,” in *3rd International Symposium on Knowledge Acquisition and Modeling (KAM)*, Wuhan, 2010.

[86] R. A. Fisher , “On the interpretation of $\chi 2$ from contingency tables, and the calculation of P,” *Journal of the Royal Statistical Society*, vol. 85, no. 1, pp. 87-94, 1922.

[87] R. A. Fisher, Statistical Methods for Research Workers, Oliver and Boyd, 1954.

[88] A. Agresti, “A Survey of Exact Inference for Contingency Tables,” *Statistical Science,* vol. 7, no. 1, pp. 131-153, 1992.

[89] H. Fazlollahtabar, H. Eslami and H. Salmani, “Designing a Fuzzy Expert System to Evaluate Alternatives in Fuzzy Analytic Hierarchy Process,” *Journal of Software Engineering & Applications,* no. 3, pp. 409-418, 2010.

[90] A. Haman and N. D. Geogranas, “Comparison of Mamdani and Sugeno Fuzzy Inference Systems for Evaluating the Quality of Experience of Hapto-Audio-Visual Applications,” in *HAVE 2008 – IEEE International Workshop on Haptic Audio Visual Environments and their Applications*, 2008.

[91] A. Kaur and A. Kaur, “Comparison of Mamdani-Type and Sugeno-Type Fuzzy Inference Systems for Air Conditioning System,” *International Journal of Soft Computing and Engineering (IJSCE)* , vol. 2, no. 2, 2012.

[92] “Fuzzy Logic Toolbox: Design and simulate fuzzy logic systems,” MathWorks,

[Online]. Available: http://www.mathworks.co.uk/products/fuzzy-logic/. [Accessed 7 December 2013].

[93] M. Cecchini, H. Aytug, et al., "Making words work: using financial text as a predictor of financial events," *Decision Support System,* vol. 50, pp. 164-175, 2010.

[94] V. Lavrenko, M. Schmill, et al., "Language models for financial news recommendation," in *International Conference on Information and Knowledge Managment,* Washington DC, 2000.

[95] M. Mittermayer, "Forecasting intraday stock price trends with text mining techniques," in *Hawaii International Conference on System Sciences* , Kailua-Kona, HI, 2004.

[96] B. Wuthrich, V. Cho, et al., "Daily stock market forecast from textual web data," in *International Conference on Systems, Man and Cybernetics,* San Diego, CA, 1998.

[97] G. Gidofalvi, C. Elkan, "Using News Articles to Predict Stock Price Movements," Department of Computer Science and Engineering, University of California, San Diego, 2003.

[98] R.P. Schumaker, Y. Zhang, et al., "Evaluating sentiment in financial news articles," *Decision Support System,* vol. 53, pp. 458-464, 2012.

[99] P. Tetlock, "Giving content to investor sentiment: the role of media in the stock market," *Journal of Finance,* vol. 62, pp. 1139-1168, 2007.

[100] P. Tetlock, M. Saar-Tsechansky et al., "More Than Words: Quantifying Language to Measure Firm's Fundamentals," in *9th Annual Texas Finance Festival*, Texas, 2007.

[101] S.R. Das, M. Y. Chen, "Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web," *Managemnet Science,* vol. 53, no. 9, pp. 1375-1388, 2007.

[102] W. Antweiller, M. Z. Frank, “Is all that talk just noise? The information content of internet stock message boards,” *The Journal of Finance LIX,* no. 3, pp. 1259-1294, 2004.

[103] A. K. Davis, J. M. Piger, L. M. Sedor, *"Beyond the Numbers: Managers' Use of Optimistic and Pessimistic Tone in Earnings Press Releases",* AAA 2008 Financial Accounting and Reporting Sections (FARS), 2008.

[104] S. Moshiri, F. Foroutan, “Forecasting nonlinear crude oil futures prices,” *Energy Journal,* vol. 27, no. 4, pp. 81-95, 2006.

[105] W. E. Shambora, R. Rossiter, “Are there exploitable inefficiencies in the futures market of oil?,” *Energy Economics,* vol. 29, no. 1, pp. 18-27, 2007.

[106] I. P. Guyon, “Applications of neural networks to character recognition,” *International Journal of Pattern Recognition and Artificial Intelligence,* vol. 5, pp. 353-382, 1991.

[107] X. Zhang, K. K. Lai, S. Wang, “A new approach for crude oil price analysis based on empirical mode decomposition,” *Energy Economics,* vol. 30, no. 3, pp. 905-918, 2008.

[108] A.J. Smola, B. Schoelkopf, “A Tutorial on Support Vector Regression∗,” 30 September 2003. [Online]. Available: alex.smola.org/papers/2003/SmoSch03b.pdf.

[109] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten, “The WEKA Data Mining Software: An Update,” *SIGKDD Explorations,* vol. 11, no. 1, 2009.

[110] A. Kumar, J. Luo, G. Bennett, “"Statistical Evaluation of Lower Flammability Distance (LFD) using Four Hazardous Release Models",” *Process Safety Progress,* vol. 12, no. 1, pp. 1-11, 1993.

[111] D. Wisell, M. Isaksson, N. Keskitalo, “A General Evaluation Criteria for

] Behavioral," in *69th ARFTG Conference*, Honolulu, 2007.

[112] M. Negnevitsky, Artificial intelligence : a guide to intelligent systems, New York : Addison-Wesley, 2005.

[113] E. Kuhn, ""CNN Politics – Political Ticker... Google unveils top political searches of 2009". CNN. Retrieved February 14, 2010.," [Online].

[114] V. J. Easton and J. H. McColl, "Statistics Glossary," [Online]. Available: http://www.stats.gla.ac.uk/steps/glossary/hypothesis_testing.html#h0. [Accessed 3 December 2013].

[115] Y. Nelson, S. Stoner, G. Gemis, &, H. Nix,, "Results of Delphi VIII," Energy Report, California Energy, California, 1994.

[116] "Organisation of Petroleum Exporting Countries (OPEC)," [Online]. Available: http://www.opec.org/opec_web/en/17.htm. [Accessed 24 July 2010].

[117] E. M. Azoff, Neural Network Time Series Forecasting of Financial Markets, John Wiley & Sons, 1994.

[118] "Investopedia," [Online]. Available: http://www.investopedia.com/terms/i/inflation.asp. [Accessed 25 October 2013].

[119] "International Monetary Fund (IMF)," [Online]. Available: www.imf.org/external/data.htm. [Accessed 18 January 2009].

[120] "International Energy Agency," *Analysis of the Impact of High Oil Prices on the Global Economy,* May 2004.

[121] "Hurricane Rita," Wikipedia, [Online]. Available: http://en.wikipedia.org/wiki/Hurricane_Rita. [Accessed 14 May 2013].

[122] "Hurricane Katrina," Wikipedia, [Online]. Available: http://en.wikipedia.org/wiki/Hurricane_Katrina. [Accessed 15 May 2013].

[123] P. Buitelaar, B. Sacaleanu, "Extending Synsets with Medical Term," in *Proceedings of the First International WordNet Conference*, Mysore, India, 2002.

[124] Z. Qu, H. Zhang, H. Li, "Determinants of Online Merchant Rating: Content Analysis of Consumer Comments about Yahoo Merchants," *Decision Support System*, vol. 46, pp. 440-449, 2008.

[125] K. He, L. Yu, K. K. Lai, "Crude oil price analysis and forecasting using wavelet decomposed ensemble model," *Energy*, vol. 46, pp. 564-574, 2012.

[126] E.F. Kelly and P.J. Stone,, "Computer Recongnition of English Word Senses,," Amsterdam:Noord-Holland.

[127] "BP Global," [Online]. Available: http://www.bp.com. [Accessed 19 June 2013].

[128] J. Philip and K. Verleger, Adjusting to Volatile Energy Prices: Policy Analyses in International Economic Series, vol. 39, Peterson Institute, 1994.

Appendix A Sample of Terminal Grammar ('Dictionary'): The *Header*

```
t
keywords
e
one
ext_0

n
coreTerm_
c
coreTerm
ext_0

n
coreTerm_
c
crudeoilprice
ext_0

n
coreTerm_
c
crudeoil
ext_0

n
coreTerm_
c
oilprice
ext_0

n
coreTerm_
c
welldrilled
ext_0

n
coreTerm_
c
provedreserves
ext_0

n
organisationpetroleumexportingcountries
ext_0

n
coreTerm_
c
organisationeconomicco-operationdevelopment
ext_0

n
crudeoilprice
c
crude#DELI;oil#DELI;price
ext_0

n
number
c
numberRegex
ext_0

t
PositiveAction
```

e
increas#DELI;rise#DELI;rose#DELI;inclin#DELI;spike#DELI;stabil#DELI;gain#DELI;remain#
DELI;up#DELI;boost#DELI;more#DELI;pump#DELI;higher#DELI;positive#DELI;add#DELI;grow#D
ELI;advanc#DELI;aggrandiz#DELI;aggravat#DELI;amplifi#DELI;annex#DELI;augment#DELI;boo
st#DELI;broad#DELI;build#DELI;deep#DELI;develop#DELI;dilat#DELI;distend#DELI;doubl#DE
LI;enhanc#DELI;enlarg#DELI;escalat#DELI;exaggerat#DELI;expand#DELI;extend#DELI;furthe
r#DELI;height#DELI;inflat#DELI;intensifi#DELI;length#DELI;magnifi#DELI;mount#DELI;mul
tipli#DELI;pad#DELI;progress#DELI;proliferat#DELI;prolong#DELI;protract#DELI;pullulat
#DELI;rais#DELI;redoubl#DELI;reinforc#DELI;ris#DELI;sharp#DELI;snowball#DELI;spread#D
ELI;strength#DELI;supplement#DELI;swarm#DELI;swell#DELI;teem#DELI;thick#DELI;triple#D
ELI;wax#DELI;wid
ext_0

t
NegativeAction
e
decreas#DELI;fall#DELI;fell#DELI;dip#DELI;declin#DELI;destabil#DELI;down#DELI;disrupt
#DELI;fallen#DELI;cut#DELI;slip#DELI;neg#DELI;decreas#DELI;grow#DELI;less#DELI;make#D
ELI;less#DELI;abat#DELI;calm#DELI;down#DELI;crumbl#DELI;curb#DELI;curtail#DELI;cut#DE
LI;decai#DELI;declin#DELI;degener#DELI;depreci#DELI;deterior#DELI;devalu#DELI;die#DEL
I;diminish#DELI;droop#DELI;drop#DELI;off#DELI;dry#DELI;dwindl#DELI;ebb#DELI;evapor#DE
LI;fade#DELI;fall#DELI;lessen#DELI;lighten#DELI;lose#DELI;edg#DELI;lower#DELI;modifi#
DELI;narrow#DELI;quell#DELI;quiet#DELI;reduc#DELI;restrain#DELI;low#DELI;settl#DELI;s
hrink#DELI;shrivel#DELI;sink#DELI;slack#DELI;slacken#DELI;slash#DELI;slump#DELI;subsi
d#DELI;tail#DELI;wane#DELI;wast#DELI;weaken#DELI;awai#DELI;wither
ext_0

t
maintainAction
e
keep#DELI;advanc#DELI;carri#DELI;conserv#DELI;continu#DELI;control#DELI;cultiv#DELI;f
inanc#DELI;keepgo#DELI;perpetu#DELI;preserv#DELI;persever#DELI;prolong#DELI;renew#DEL
I;retain#DELI;support#DELI;sustain#DELI;uphold
ext_0

t
coreTerm
e
demand#DELI;consumpt#DELI;suppli#DELI;product#DELI;inventori#DELI;well#DELI;reserv#DE
LI;stock#DELI;import#DELI;petroleum#DELI;refineri#DELI;spill#DELI;leak#DELI;barrel#DE
LI;usd/barrel#DELI;opec#DELI;oecd#DELI;nymex#DELI;output
ext_0

t
crude
e
crude
ext_0

t
oil
e
oil
ext_0

t
price
e
price#DELI;prices
ext_0

t
well
e
well
ext_0

n
importCountries_
c
importCountries
ext_0

```
n
Date
c
month#DELI;day
ext_0

t
day
r
[0-9]|1[0-9]|2[0-9]|(3[01])
ext_1
x
1~1

t
dateRegex
r
(0[0-9]|1[0-9]|2[0-9]|3[01])(-|/)(0[0-9]|1[0-9]|2[0-9]|3[01])(-|/)((19|20)*[0-9][0-
9])
ext_0

t
month
e
january#DELI;february#DELI;march#DELI;april#DELI;may#DELI;june#DELI;july#DELI;august#
DELI;september#DELI;october#DELI;november#DELI;december#DELI;jan#DELI;feb#DELI;mar#DE
LI;apr#DELI;may#DELI;june#DELI;jul#DELI;aug#DELI;sept#DELI;oct#DELI;nov#DELI;dec
ext_0

t
year
r
(19|20)*[0-9][0-9]
ext_1
x
2008~1
```

Appendix B Sample of Fuzzy Inference System Structure

```
[System]
Name='mamdani2'
Type='mamdani'
Version=2.0
NumInputs=10
NumOutputs=1
NumRules=25
AndMethod='min'
OrMethod='max'
ImpMethod='min'
AggMethod='max'
DefuzzMethod='centroid'

[Input1]
Name='recession'
Range=[-1 1]
NumMFs=3
MF1='negative':'trapmf',[-1.72 -1.08 -0.92 -0.28]
MF2='neutral':'trimf',[-0.8 0 0.8]
MF3='positive':'trapmf',[0.28 0.92 1.08 1.72]

[Input2]
Name='demand'
Range=[-1 1]
NumMFs=3
MF1='negative':'trapmf',[-1.72 -1.08 -0.92 -0.28]
MF2='neutral':'trimf',[-0.8 0 0.8]
MF3='positive':'trapmf',[0.28 0.92 1.08 1.72]

[Input3]
Name='refinery'
Range=[-1 1]
NumMFs=3
MF1='negative':'trapmf',[-1.72 -1.08 -0.92 -0.28]
MF2='neutral':'trimf',[-0.8 0 0.8]
MF3='positive':'trapmf',[0.28 0.92 1.08 1.72]

[Input4]
Name='stocks'
Range=[-1 1]
NumMFs=3
MF1='negative':'trapmf',[-1.72 -1.08 -0.92 -0.28]
MF2='neutral':'trimf',[-0.8 0 0.8]
MF3='positive':'trapmf',[0.28 0.92 1.08 1.72]

[Input5]
Name='supply'
Range=[-1 1]
NumMFs=3
MF1='negative':'trapmf',[-1.72 -1.08 -0.92 -0.28]
MF2='neutral':'trimf',[-0.8 0 0.8]
MF3='positive':'trapmf',[0.28 0.92 1.08 1.72]
```

```
[Input6]
Name='war'
Range=[-1 1]
NumMFs=3
MF1='negative':'trapmf',[-1.72 -1.08 -0.92 -0.28]
MF2='neutral':'trimf',[-0.8 0 0.8]
MF3='positive':'trapmf',[0.28 0.92 1.08 1.72]

[Input7]
Name='opec_decision'
Range=[-1 1]
NumMFs=3
MF1='negative':'trapmf',[-1.72 -1.08 -0.92 -0.28]
MF2='neutral':'trimf',[-0.8 0 0.8]
MF3='positive':'trapmf',[0.28 0.92 1.08 1.72]

[Input8]
Name='consumption'
Range=[-1 1]
NumMFs=3
MF1='negative':'trapmf',[-1.72 -1.08 -0.92 -0.28]
MF2='neutral':'trimf',[-0.8 0 0.8]
MF3='positive':'trapmf',[0.28 0.92 1.08 1.72]

[Input9]
Name='production'
Range=[-1 1]
NumMFs=3
MF1='negative':'trapmf',[-1.72 -1.08 -0.92 -0.28]
MF2='neutral':'trimf',[-0.8 0 0.8]
MF3='positive':'trapmf',[0.28 0.92 1.08 1.72]

[Input10]
Name='import'
Range=[-1 1]
NumMFs=3
MF1='negative':'trapmf',[-1.72 -1.08 -0.92 -0.28]
MF2='neutral':'trimf',[-0.8 0 0.8]
MF3='positive':'trimf',[0.2 1 1.8]

[Output1]
Name='price'
Range=[0 1]
NumMFs=2
MF1='negative':'trimf',[-1 0 1]
MF2='positive':'trimf',[0 1 2]
```

Appendix C Linguistic Prediction Model: The Computation

This section lists the samples of computations made for performance evaluation in lingusitic model. The calculations include the computation of D_{stat}, RMSE, MSE, and NMSE with the application of Microsoft Excels program.

A-BEST PREDICTION RESULTS BASED ON TESTING DATA WITH 90:10 RATIO

Table A-1 Best Prediction Results For Directional Price with Hyperbolic Tangent, Hidden Layer Neuron=5

Hyperbolic Tangent: HL5									
Previous Month Value	No.	Ori (Y)	Output Result_1 (Y')	Error (Y-Y')	Squared Error (Y-Y')^2	Average Y' (A)	(Y'-A)^2	Dstat (Y1-Y0)(Y'1-Y0)	
1	1	-1	0.9704	-1.9704	3.8823	0.6300	0.1159	0.0593	1
	2	-1	0.2896	-1.2896	1.6630	0.6300	0.1159	0.0000	1
Total					**5.54533**		**0.23173**		**2**
Average $\hat{Y}$ (A)									**0.62997**
MSE: Total $(Y-\hat{Y})^2$ / Total (N)									**2.77267**
RMSE: √MSE									**1.66513**
NMSE: MSE /Total $[(\hat{Y}-A)^2/N]$									**...**
Dstat %: (Y1-Y0)(Y'1-Y0), Total(True)/N									**100**

Table A-2 Best Prediction Results For Disnormalised Price with Hyperbolic Tangent, Hidden Layer Neuron=4

Hyperbolic Tangent: HL4 (Disnormalised)									
Previous Month Value	No.	Ori Y	Output Result_1 Y'	Error (Y-Y')	Squared Error (Y-Y')^2	Average Y' (A)	(Y'-A)^2	Dstat (Y1-Y0)(Y'1-Y0)	
75.82	1	78.08	89.0153	-10.9353	119.5816	71.5448	305.2204	29.8215	1
	2	74.3	54.0742	20.2258	409.0822	71.5448	305.2204	4017.7146	1
Total					**528.6637**		**610.4408**		**1**
Average $\hat{Y}$ (A)									**71.54478**
MSE: Total $(Y-\hat{Y})^2$ / Total (N)									**264.3319**
RMSE: √MSE									**...**
NMSE: MSE /Total $[(\hat{Y}-A)^2/N]$									**0.866036**
Dstat %: (Y1-Y0)(Y'1-Y0), Total(True)/N									**100**

Table A-3 Best Prediction Results For Original Price with Log Sigmoid, Hidden Layer Neuron=4

Log-Sigmoid: HL4									
Previous Month Value	No.	Ori Y	Output Result_1 Y'	Error (Y-Y')	Squared Error (Y-Y')^2	Average Y' (A)	(Y'-A)^2	Dstat (Y1-Y0)(Y'1-Y0)	
75.82	1	78.08	47.0382	31.0418	963.5930	64.6860	311.4460	-65.0469	0
	2	74.30	82.3339	-8.0339	64.5431	64.6860	311.4460	-16.0796	0
Total					**1028.1361**		**622.8920**		**0**
Average $\hat{Y}$ (A)									**64.6860**
MSE: Total $(Y-\hat{Y})^2$ / Total (N)									**514.0681**
RMSE: √MSE									**22.6731**
NMSE: MSE /Total $[(\hat{Y}-A)^2/N]$									**...**
Dstat %: (Y1-Y0)(Y'1-Y0), Total(True)/N									**0**

B- BEST PREDICTION RESULTS BASED ON TESTING DATA WITH 80:20 RATIO

Table B-1 Best Prediction Results For Directional Price with Hyperbolic Tangent, Hidden Layer Neuron=4

Hyperbolic Tangent: HL4									
Previous Month	No. (N)	Ori (Y)	Output Result_1 (Y')	Error (Y-Y')	Squared Error (Y-Y')^2	Average Y' (A)	(Y'-A)^2	Dstat (Y1-Y0)(Y'1-Y0)	
-1	1	-1	-0.85691	-0.14309	0.02048	0.26869	1.26697	0.00000	1
	2	1	1.11098	-0.11098	0.01232	0.26869	0.70946	4.22197	1
	3	1	1.08698	-0.08698	0.00756	0.26869	0.66959	0.00000	1
	4	-1	1.09710	-2.09710	4.39781	0.26869	0.68626	-0.19419	0
	5	-1	-1.09470	0.09470	0.00897	0.26869	1.85883	0.00000	1
Total					**0.03279**		**1.97643**		**4**
Average $\hat{Y}$ (A)									**0.26869**
MSE: Total $(Y-\hat{Y})^2$ / Total (N)									**0.00656**
RMSE: √MSE									**0.08099**
NMSE: MSE /Total $[(\hat{Y}-A)^2/N]$									**0.01659**
Dstat %: (Y1-Y0)(Y'1-Y0), Total(True)/N									**80**

Table B-2 Best Prediction Results For Normalised Price with Hyperbolic Tangent, Hidden Layer Neuron=5

Hyperbolic Tangent: HL5									
Previous Month	No. (N)	Ori (Y)	Output Result_1 (Y')	Error (Y-Y')	Squared Error (Y-Y')^2	Average Y' (A)	(Y'-A)^2	Dstat (Y1-Y0)(Y'1-Y0)	
-0.09572	1	0.12776	0.09282	0.03494	0.00122	0.04931	0.00189	0.04214	1
	2	-0.02252	0.08396	-0.10648	0.01134	0.04931	0.00120	0.00658	1
	3	0.09156	0.16984	-0.07827	0.00613	0.04931	0.01453	0.02194	1
	4	0.02981	0.20729	-0.17748	0.03150	0.04931	0.02496	-0.00715	0
	5	-0.04841	-0.30737	0.25896	0.06706	0.04931	0.12722	0.02637	1
Total					**0.11725**		**0.1698**		**4**
Average $\hat{Y}$ (A)									**0.04931**
MSE: Total $(Y-\hat{Y})^2$ / Total (N)									**0.02345**
RMSE: √MSE									**0.15313**
NMSE: MSE /Total $[(\hat{Y}-A)^2/N]$									**0.69049**
Dstat %: (Y1-Y0)(Y'1-Y0), Total(True)/N									**80**

Table B-3 Best Prediction Results For Original Price with Hyperbolic Tangent, Hidden Layer Neuron=5

Hyperbolic Tangent: HL5									
Previous Month	No. (N)	Ori (Y)	Output Result_1 (Y')	Error (Y-Y')	Squared Error (Y-Y')^2	Average Y' (A)	(Y'-A)^2	Dstat (Y1-Y0)(Y'1-Y0)	
63.010	1	71.06	37.68298	33.37702	1114.02552	74.41384	1349.15597	-203.88252	0
	2	69.46	56.73392	12.72608	161.95313	74.41384	312.57952	22.92173	1
	3	75.82	98.38144	-22.56144	509.01847	74.41384	574.44584	183.94034	1
	4	78.08	55.69544	22.38456	501.06850	74.41384	350.37839	-45.48150	0
	5	74.30	123.57541	-49.27541	2428.06626	74.41384	2416.86041	-171.97266	0
Total					**4714.13188**		**5003.42014**		**2**
Average $\hat{Y}$ (A)									**74.41384**
MSE: Total $(Y- \hat{Y})^2$ / Total (N)									**942.82638**
RMSE: √MSE									**30.70548**
NMSE: MSE /Total $[(\hat{Y}-A)^2/ N]$									**0.94218**
Dstat %: (Y1-Y0)(Y'1-Y0), Total(True)/N									**40**

C- BEST PREDICTION RESULTS BASED ON TESTING DATA WITH 70:30 RATIO

Table C-1 Best Prediction Results For Directional Price with Hyperbolic Tangent, Hidden Layer Neuron=5

Hyperbolic Tangent: HL5									
Previous Month	No. (N)	Ori (Y)	Output Result_1 (Y')	Error (Y-Y')	Squared Error (Y-Y')^2	Average Y' (A)	(Y'-A)^2	Dstat (Y1-Y0)(Y'1-Y0)	
1	1	1	1.0438	-0.04385	0.00192	0.73466	0.09560	0.00000	1
	2	-1	1.0493	-2.04929	4.19961	0.73466	0.09900	-0.09859	0
	3	-1	1.0058	-2.00581	4.02326	0.73466	0.07352	0.00000	1
	4	1	1.0398	-0.03984	0.00159	0.73466	0.09314	4.07968	1
	5	1	1.0519	-0.05186	0.00269	0.73466	0.10061	0.00000	1
	6	-1	1.0453	-2.04534	4.18340	0.73466	0.09652	-0.09067	0
	7	-1	-1.0934	0.09337	0.00872	0.73466	3.34169	0.00000	1
Total					**12.42119**		**3.90007**		**5**
Average $\hat{Y}$ (A)									**0.73466**
MSE: Total $(Y-\hat{Y})^2$ / Total (N)									**1.77446**
RMSE: √MSE									**1.33209**
NMSE: MSE /Total $[(\hat{Y}-A)^2/N]$									**3.18486**
Dstat %: (Y1-Y0)(Y'1-Y0), Total(True)/N									**71.42857**

Table C-2 Best Prediction Results For Normalised Price with Hyperbolic Tangent, Hidden Layer Neuron=5

Hyperbolic Tangent: HL5									
Previous Month	No. (N)	Ori (Y)	Output Result_1 (Y')	Error (Y-Y')	Squared Error (Y-Y')^2	Average Y' (A)	(Y'-A)^2	Dstat (Y1-Y0)(Y'1-Y0)	
0.18819	1	0.177823	0.23758	-0.05976	0.00357	0.11140	0.01592	-0.00051	0
	2	-0.095723	0.12440	-0.22013	0.04846	0.11140	0.00017	0.01461	1
	3	0.127757	0.09365	0.03411	0.00116	0.11140	0.00032	0.04232	1
	4	-0.022516	0.15783	-0.18035	0.03253	0.11140	0.00216	-0.00452	0
	5	0.091563	0.23703	-0.14546	0.02116	0.11140	0.01578	0.02961	1
	6	0.029807	0.23164	-0.20183	0.04073	0.11140	0.01446	-0.00865	0
	7	-0.048412	-0.30231	0.25390	0.06446	0.11140	0.17116	0.02598	1
Total					**0.21207**		**0.21996**		**4**
Average $\hat{Y}$ (A)									**0.11140**
MSE: Total $(Y-\hat{Y})^2$ / Total (N)									**0.03030**
RMSE: √MSE									**0.17406**
NMSE: MSE /Total $[(\hat{Y}-A)^2/N]$									**0.96417**
Dstat %: (Y1-Y0)(Y'1-Y0), Total(True)/N									**57.14286**

Table C-3 Best Prediction Results For Original Price with Hyperbolic Tangent, Hidden Layer Neuron=5

Hyperbolic Tangent: HL5									
Previous Month	No. (N)	Ori (Y)	Output Result_1 (Y')	Error (Y-Y')	Squared Error (Y-Y')^2	Average Y' (A)	(Y'-A)^2	Dstat (Y1-Y0)(Y'1-Y0)	
59.160	1	69.68	82.39630	-12.71630	161.70423	81.76470	0.39891	244.44585	1
	2	63.01	85.65684	-22.64684	512.87925	81.76470	15.14870	-106.56551	0
	3	71.06	35.14933	35.91067	1289.57604	81.76470	2172.99293	-224.27837	0
	4	69.46	133.11482	-63.65482	4051.93606	81.76470	2636.83430	-99.28771	0
	5	75.82	106.11000	-30.29000	917.48436	81.76470	592.69361	233.09403	1
	6	78.08	52.92775	25.15225	632.63589	81.76470	831.57019	-51.73649	0
	7	74.30	76.99790	-2.69790	7.27864	81.76470	22.72247	4.09035	1
Total					**7573.49446**		**6272.36110**		**3**
Average $\hat{Y}$ (A)									**81.76470**
MSE: Total $(Y-\hat{Y})^2$ / Total (N)									**1081.92778**
RMSE: √MSE									**32.89267**
NMSE: MSE /Total $[(\hat{Y}-A)^2/N]$									**1.20744**
Dstat %: (Y1-Y0)(Y'1-Y0), Total(True)/N									**42.85714**