

**A Big Data Study on the Sentiment Analysis of Social Networks and Nonlinear System Modelling**

**By:**

Youchen Wang (130233192)

The University of Sheffield

Faculty of Engineering

Department of Automatic Control and Systems Engineering

Submission Date: 25/01/2018

# Acknowledgements

I would like to express my most sincere gratitude to all the people who have helped me during my PhD study.

Please allow me to express my special appreciation and thanks to my supervisor Dr. Hua-Liang Wei who has been a tremendous mentor to me. During my PhD study, I had many difficulties in academic problems and thanks again for your patience and encouragement. Your advice on both research as well as on my career has been priceless. Similarly, profound gratitude goes to my second supervisors: Prof. Robert F Harrison and Prof. Qing-Chang Zhong.

I would also like to thank my group members Dr. Fei He and Jia Zhao, who have truly been my academic mentors. I express my particularly thanks to Jia Zhao for his generous guidance in my modelling work, and for hosting me in many restaurants in Sheffield. I have very fond memories of my time there.

Special thanks to my father, Hairong Wang and my mother, Hong Ma. Thanks for your support during my PhD study, I do not know if I would have been able to have completed my PhD research without your support. Thank you to my parents for all the sacrifices that you have made for me. Finally, I would like express sincerely appreciation to my beloved girlfriend Jiaqi Wei who has always supported me in the background. They are the most important people in my life and I dedicate this thesis to them.

# Abstract

A Big Data Study on the Sentiment Analysis of Social Networks and Nonlinear System Modelling

Youchen Wang

In the big data age, the development of social network services has already changed people’s way of life. Twitter, as one of the most popular microblogging service, has profoundly influenced and changed our daily life. Twitter users are discussing different kinds of topics, include celebrities, movies, economics, the military and politics. Considering about the number of Twitter users, Twitter may contain numerous useful information. Based on the behaviour psychology, these rich-sentiment data can easily affect other people, especially in consumption behaviour, investment and political issue. Therefore, extract and analysis of Twitter interactive data may help researchers to investigate the political issues and economic systems.

This thesis introduces an original programme based on Twitter API and R programming language. This programme applied Twitter keywords search function to obtain related tweets, these opinion-rich datasets about tweets contents, tweets’ authors and tweets post time on Twitter can be extracted by Twitter API and R programming. In order to collect more comprehensive Twitter sentiment about political and economic issues, this programme has been extended to geography location search and post time search. This Twitter data extracting method is widely applied in this thesis: there are over 3 million tweets about 2016 US presidential election; 23332 tweets about 2016 UK Brexit referendum; around 90000 daily tweets related to FTSE100 are extracted.

A novel text pre-process method for Twitter data is proposed and discussed. The extracted tweets may contain a variety of interference information such as different languages, links, @ someone and garbled. The text pre-process method includes: keep English Twitter and filter other languages Twitter; get the frequency of key sentiment words; reduce interference from garbled, links and @ someone.

The NRC lexicon for sentiment analysis has been utilized to real world problems to explore: Twitter sentiment and emotion index daily change about Hillary Clinton and Donald Trump during the period of US presidential election; Twitter sentiment in different parts of UK towards Brexit referendum; daily Twitter sentiment index about UK stock market. According to these datasets, we investigate whether the collective sentiment on Twitter can help to visualize, model and predict these political issues.

For the first time, this thesis proposed a hybrid model for Twitter sentiment classification. A novel feature selection methods based on NRC lexicon and classic classification algorithms KNN and Naïve Bayes are combined to improve the performance of Twitter polarity classification. The results are evaluated and validated.

Furthermore, this thesis employed wavelet based nonlinear models on stock market systems. There are two case studies has been discussed: the first one is about crude oil price and FTSE100 system; the second one on the study of Twitter sentiment & FTSE100 system. Although applying crude oil price and Twitter sentiment index to model stock market change has been studied by Granger Causality test and ANN related algorithms, this thesis firstly using Wavelet based NARX to model these processes.

Keywords-component; Sentiment Analysis; Lexicon Based method; Machine Learning Twitter; Wavelet Models, Brexit, US Presidential, FTSE.

# Acronyms

maximum lag for input time series

maximum lag for output time series

Akaike Information Criterion

ANN Artificial Neural Network

AR AutoRegression Model

ARMA AutoRegression Moving Average Model

ARMAX AutoRegression Moving Average with eXogenous inputs Model

ARX AutoRegression with eXogenous inputs Model

Bayesian Information Criterion

Cross Validation

CWT Continuous Wavelet Transform

DWT Discrete Wavelet Transform

EEG Electroencephalography

EMG Electromyography

ERR Error Reduction Ratio

IE Information Retrieval

IR Information Extraction

GC Granger Causality test

KNN K-Nearest Neighbors algorithm

MAE Mean Absolute Error

MSE Mean Square Error

NB Naïve Bayes algorithm

NRC NRC Sentiment Lexicon

NARMAX Nonlinear AutoRegressive Moving Average with eXogenous inputs Model

NARX Nonlinear AutoRegressive with eXogenous inputs Model

OLS Orthogonal Least Squares

RMSE Root Mean Square Error

SNA Social Network Analysis

SVM Support Vector Machine

TF Term Frequency

IDF Inverse Document Frequency

TF-IDF Term Frequency-Inverse Document Frequency

VAR Vector AutoRegression

WMISO Wavelet Multi-Input Single-Output System

DJIA Dow Jones Industrial Average

Table of Contents

[Acknowledgements 2](#_Toc502566395)

[Abstract 3](#_Toc502566396)

[Acronyms 5](#_Toc502566397)

[List of Tables 13](#_Toc502566398)

[List of Figures 15](#_Toc502566399)

[Chapter 1. Introduction 20](#_Toc502566400)

[1.1. Background 20](#_Toc502566401)

[1.2. Motivation 21](#_Toc502566402)

[1.3. Overview 23](#_Toc502566403)

[1.4. Contributions 25](#_Toc502566404)

[Chapter 2. Literature Review 28](#_Toc502566405)

[2.1. Introduction 28](#_Toc502566406)

[2.2. Modelling and Forecasting Methods 29](#_Toc502566407)

[2.2.1. Introduction 29](#_Toc502566408)

[2.2.2. Linear Models 29](#_Toc502566409)

[2.2.3. Nonlinear Models 30](#_Toc502566410)

[2.2.4. Granger Causality Test 31](#_Toc502566411)

[2.2.5. Artificial Neural Network 33](#_Toc502566412)

[2.2.6. Wavelet Pre-process for Nonlinear System Identifications 36](#_Toc502566413)

[2.3. The Influence of Twitter Sentiment 41](#_Toc502566414)

[2.3.1. Background 41](#_Toc502566415)

[2.3.2. What Makes Twitter Sentiment Significant 43](#_Toc502566416)

[2.3.3. Twitter Network Communication Analysis 45](#_Toc502566417)

[2.3.4. Web Mining 47](#_Toc502566418)

[2.3.5. How to Extract Tweets on Twitter 48](#_Toc502566419)

[2.3.6. Web Mining and Twitter Sentiment Applications 54](#_Toc502566420)

[2.3.7. Twitter Sentiment Influence on Political Election 61](#_Toc502566421)

[2.3.8. Twitter Sentiment Influence on Stock Market Index 62](#_Toc502566422)

[2.3.9. Twitter Sentiment Influence on Brexit 63](#_Toc502566423)

[2.3.10. The Application of Twitter Sentiment Analysis 64](#_Toc502566424)

[2.4. Sentiment Analysis Methods 65](#_Toc502566425)

[2.4.1. Background and Introduction 65](#_Toc502566426)

[2.4.2. Twitter data Pre-process 67](#_Toc502566427)

[2.4.3. Lexicon Based Method 68](#_Toc502566428)

[2.4.4. Text Mining 68](#_Toc502566429)

[2.4.5. Machine Learning Methods for Document Classification 69](#_Toc502566430)

[2.4.6. How the Machine Learning Algorithm Affects this Research? 74](#_Toc502566431)

[2.5. Social Networks and Complex Network 75](#_Toc502566432)

[2.5.1. Introduction 75](#_Toc502566433)

[2.5.2. Complex Network 76](#_Toc502566434)

[2.5.3. Complex Network Properties 76](#_Toc502566435)

[2.5.4. Social Network 77](#_Toc502566436)

[2.5.5. Complex/Social Network Platform 79](#_Toc502566437)

[2.6. Conclusion 81](#_Toc502566438)

[Chapter 3. Sentiment Analysis for Web Information 85](#_Toc502566439)

[3.1. Introduction 85](#_Toc502566440)

[3.2. The Significance of Twitter Information 86](#_Toc502566441)

[3.3. How to Extract Tweets on Twitter 87](#_Toc502566442)

[3.3.1. Twitter extraction with R 87](#_Toc502566443)

[3.3.2. FTSE Twitter Word Cloud 88](#_Toc502566444)

[3.4. Twitter Data Pre-process 89](#_Toc502566445)

[3.5. Sentiment Analysis for Twitter 94](#_Toc502566446)

[3.5.1. Introduction 94](#_Toc502566447)

[3.5.2. Twitter Sentiment Analysis about Hillary Clinton and Donald Trump 95](#_Toc502566448)

[3.5.3. Twitter Emotion Analysis about Hillary Clinton and Donald Trump 99](#_Toc502566449)

[3.6. Twitter Sentiment for Brexit 2016 106](#_Toc502566450)

[3.6.1. Introduction 107](#_Toc502566451)

[3.6.2. Lexicon based method NRC 107](#_Toc502566452)

[3.6.3. Results Analysis 111](#_Toc502566453)

[3.7. Twitter Sentiment for UK stock market 111](#_Toc502566454)

[3.7.1. Background 111](#_Toc502566455)

[3.7.2. Data preparation 112](#_Toc502566456)

[3.7.3. Lexicon based method 112](#_Toc502566457)

[3.8. Conclusion 115](#_Toc502566458)

[Chapter 4 Machine Learning on Sentiment Analysis and Complex Network 117](#_Toc502566459)

[4.1. Introduction and Background 117](#_Toc502566460)

[4.2. Twitter Data Pre-process 118](#_Toc502566461)

[4.3. Feature Selection for Twitter Data 118](#_Toc502566462)

[4.3.1. Traditional Feature Selection Methods 118](#_Toc502566463)

[4.3.2. Feature selection based on NRC lexicon 119](#_Toc502566464)

[4.4. The Research on Text Classification Algorithm 121](#_Toc502566465)

[4.4.1. Naïve Bayes Classifier 121](#_Toc502566466)

[4.4.2. KNN Classifier 121](#_Toc502566467)

[4.5. NRC based Machine Learning Methods on Twitter Sentiment Analysis 122](#_Toc502566468)

[4.5.1. Experiment Background 123](#_Toc502566469)

[4.5.2. NRC based KNN Classifier 123](#_Toc502566470)

[4.5.3. NRC based Naïve Bayes (NB) Classifier 124](#_Toc502566471)

[4.5.4. NRC based KNN and Naïve Bayes Classifier Result Analysis 126](#_Toc502566472)

[4.6. Twitter Social Network Analysis 128](#_Toc502566473)

[4.6.1. Data Resources 128](#_Toc502566474)

[4.6.2. Analysis 128](#_Toc502566475)

[4.6.3. Summary 129](#_Toc502566476)

[4.7. Conclusion 130](#_Toc502566477)

[Chapter 5. Stock Market System Modelling – Wavelet Regression Model 131](#_Toc502566478)

[5.1 Introduction 131](#_Toc502566479)

[5.2 Shang Hai Composite (SSE) Index Model Representation 132](#_Toc502566480)

[5.3 Wavelet Analysis 133](#_Toc502566481)

[5.3.1 Wavelet background 133](#_Toc502566482)

[5.3.2 Wavelet transforms 134](#_Toc502566483)

[5.3.3 Selection of Mother Wavelet Function 135](#_Toc502566484)

[5.3.4 Stock Market Data Pre-process Using Discrete Wavelet Transform (DWT) 136](#_Toc502566485)

[5.4 Linear Wavelet Multi Input Single Output (WMISO) Model 140](#_Toc502566486)

[5.4.1 WMISO Model Framework 140](#_Toc502566487)

[5.4.2 Selection of Input Variables 141](#_Toc502566488)

[5.4.3 Wavelet ARX and Wavelet ARMAX 143](#_Toc502566489)

[5.4.4 Model Structure and Results Analysis 143](#_Toc502566490)

[5.5 Nonlinear Wavelet Model 146](#_Toc502566491)

[5.5.1 Orthogonal Least Square Method 147](#_Toc502566492)

[5.5.2 Model Validation 149](#_Toc502566493)

[5.6 Crude Oil price & FTSE100 Wavelet Model 152](#_Toc502566494)

[5.7 Twitter Sentiment and Twitter Emotion Predict Stock Market 158](#_Toc502566495)

[Chapter 6. Conclusion 172](#_Toc502566496)

[References 176](#_Toc502566497)

# List of Tables

[Table 2. 1 Relationship of classification evaluation 73](#_Toc502566546)

[Table 3. 1 Statistics of Trump and Hillary Popularity 97](#_Toc502566547)

[Table 3. 2 Statistics of Trump and Hillary 97](#_Toc502566548)

[Table 3. 3 Popularity of Hillary and Trump on some important dates 97](#_Toc502566549)

[Table 3. 4 Twitter emotion distribution by days 106](#_Toc502566550)

[Table 3. 5 Twitter sentiment results in central UK 108](#_Toc502566551)

[Table 3. 6 Twitter sentiment result in south UK 109](#_Toc502566552)

[Table 3. 7 Twitter sentiment result in north UK 111](#_Toc502566553)

[Figure 4. 1 Donald Trump Twitter emotion distribution 120](#_Toc502566554)

[Figure 4. 2 KNN Classification Process 122](#_Toc502566555)

[Figure 4. 3 The process of NRC based KNN classifier 124](#_Toc502566556)

[Figure 4. 4 The process of NRC based KNN classifier 125](#_Toc502566557)

[Figure 4. 5 Social network Twitter sentiment about FTSE100 in 18/11/2014 129](#_Toc502566558)

[Table 4. 1 The performance of NRC KNN classifier 127](#_Toc502566559)

[Table 4. 2 The performance of NRC NB classifier 127](#_Toc502566560)

[Table 5. 1 Cross correlation analysis about DWT FTSE 100 index and SEE composite index 141](#_Toc502566561)

[Table 5. 2 Cross correlation analysis about DWT HangSheng index and SEE composite index 141](#_Toc502566562)

[Table 5. 3 Cross correlation analysis about DWT DAX index and SEE composite index 141](#_Toc502566563)

[Table 5. 4 Cross correlation analysis about DWT CAC index and SEE composite index 142](#_Toc502566564)

[Table 5. 5 Cross correlation analysis about SP500 index and SEE composite index 142](#_Toc502566565)

[Table 5. 6 One day ahead prediction of WARX and WARMAX model on SSE composite index 145](#_Toc502566566)

[Table 5. 7 Identification of SEE system 149](#_Toc502566567)

[Table 5. 8 Model performance for SEE system 152](#_Toc502566568)

[Table 5. 9 Identification of Daily FTSE OP system 155](#_Toc502566569)

[Table 5. 10 Identification of weekly FTSE OP system 156](#_Toc502566570)

[Table 5. 11 The performance of Wavelet NARX and NARX about Twitter FTSE system 171](#_Toc502566571)

# List of Figures

[Figure 2. 1 Feedback ANN architecture (Kantardzic, 2011) 34](#_Toc502566498)

[Figure 2. 2 Recurrent ANN architecture (Kantardzic, 2011) 34](#_Toc502566499)

[Figure 2. 3 Simple Twitter dissemination process 45](#_Toc502566500)

[Figure 2. 4 Web Mining Systematics 48](#_Toc502566501)

[Figure 2. 5 Google Spread Sheet for Twitter Extraction 51](#_Toc502566502)

[Figure 2. 6 Webharvey Operator Interface 52](#_Toc502566503)

[Figure 2. 7 Webharvey Miner Data 53](#_Toc502566504)

[Figure 2. 8 Twitter API 54](#_Toc502566505)

[Figure 2. 9 Flow chart of Web data mining 56](#_Toc502566506)

[Figure 3. 1 Retrieving Tweets Results 87](#_Toc502564480)

[Figure 3. 2 FTSE Word Cloud 88](#_Toc502564481)

[Figure 3. 3 Unprocessed Tweets 90](#_Toc502564482)

[Figure 3. 4 Pre-Processed Tweets 91](#_Toc502564483)

[Figure 3. 5 Sample Donald Trump’s Tweets 92](#_Toc502564484)

[Figure 3. 6 Sample Twitter Word Frequency of Donald Trump 93](#_Toc502564485)

[Figure 3. 7 Sample Twitter Word Cloud of Donald Trump 94](#_Toc502564486)

[Figure 3. 8 Daily sentiment index change of Hillary Clinton 96](#_Toc502564487)

[Figure 3. 9 Daily sentiment index change about Donald Trump 96](#_Toc502564488)

[Figure 3. 10 Difference between Clinton and Trump positive Twitter sentiment index 98](#_Toc502564489)

[Figure 3. 11 Difference between Clinton and Trump Negative Twitter sentiment index 99](#_Toc502564490)

[Figure 3. 12 Daily Emotion index about Hillary Clinton 100](#_Toc502564491)

[Figure 3. 13 Daily emotion index about Donald Trump 100](#_Toc502564492)

[Figure 3. 14 Difference of Twitter anger emotion time series about Hillary and Trump 101](#_Toc502564493)

[Figure 3. 15 Difference of Twitter anticipation emotion time series about Hillary and Trump 102](#_Toc502564494)

[Figure 3. 16 Difference of Twitter disgust emotion time series about Hillary and Trump 102](#_Toc502564495)

[Figure 3. 17 Difference of Twitter fear emotion time series about Hillary and Trump 103](#_Toc502564496)

[Figure 3. 18 Difference of Twitter joy emotion time series about Hillary and Trump 103](#_Toc502564497)

[Figure 3. 19 Difference of Twitter sadness emotion time series about Hillary and Trump 104](#_Toc502564498)

[Figure 3. 20 Difference of Twitter surprise emotion time series about Hillary and Trump 104](#_Toc502564499)

[Figure 3. 21 Difference of Twitter trust emotion time series about Hillary and Trump 105](#_Toc502564500)

[Figure 3. 22 Twitter emotion index comparison between Hillary and Trump 106](#_Toc502564501)

[Figure 3. 23 Twitter Sentiment about Brexit in the central UK 108](#_Toc502564502)

[Figure 3. 24 Brexit Twitter Sentiment in London Area 109](#_Toc502564503)

[Figure 3. 25 Brexit Twitter sentiment in North UK 110](#_Toc502564504)

[Figure 3. 26 FTSE Twitter sentiment index 113](#_Toc502564505)

[Figure 3. 27 Twitter polar index bar chart 114](#_Toc502564506)

[Figure 3. 28 FTSE Twitter Emotion Index 114](#_Toc502564507)

[Figure 4. 1 Donald Trump Twitter emotion distribution 120](#_Toc502566507)

[Figure 4. 2 KNN Classification Process 122](#_Toc502566508)

[Figure 4. 3 The process of NRC based KNN classifier 124](#_Toc502566509)

[Figure 4. 4 The process of NRC based KNN classifier 125](#_Toc502566510)

[Figure 4. 5 Social network Twitter sentiment about FTSE100 in 18/11/2014 129](#_Toc502566511)

[Figure 5. 1 Wavelet Decomposition of FTSE 100 index time series 137](#_Toc502566512)

[Figure 5. 2 Wavelet Decomposition of SEE Composite index time series 137](#_Toc502566513)

[Figure 5. 3 Wavelet Decomposition of HangSheng index time series 138](#_Toc502566514)

[Figure 5. 4 Wavelet Decomposition of DAX index time series 138](#_Toc502566515)

[Figure 5. 5 Wavelet Decomposition of CAC index time series 139](#_Toc502566516)

[Figure 5. 6 Wavelet Decomposition of SP500 index time series 139](#_Toc502566517)

[Figure 5. 7 WMISO Model Structure 140](#_Toc502566518)

[Figure 5. 8 Wavelet linear regression model framework 143](#_Toc502566519)

[Figure 5. 9 WARX and WARMAX training model result 144](#_Toc502566520)

[Figure 5. 10 WARX and WARMAX validation model result 145](#_Toc502566521)

[Figure 5. 11 Nonlinear Wavelet Model Structure 146](#_Toc502566522)

[Figure 5. 12 Simulation result of training data 151](#_Toc502566523)

[Figure 5. 13 Simulation results of validation data 151](#_Toc502566524)

[Figure 5. 14 Nonlinear wavelet model structure 154](#_Toc502566525)

[Figure 5. 15 Simulation results of daily FTSE & OP model 156](#_Toc502566526)

[Figure 5. 16 Simulation results of weekly FTSE & OP model validation 157](#_Toc502566527)

[Figure 5. 17 Wavelet Decomposition of Twitter positive index 159](#_Toc502566528)

[Figure 5. 18 Wavelet Decomposition of Twitter negative index 159](#_Toc502566529)

[Figure 5. 19 Wavelet Decomposition of Twitter anger index 160](#_Toc502566530)

[Figure 5. 20 Wavelet Decomposition of Twitter anticipation index 160](#_Toc502566531)

[Figure 5. 21 Wavelet Decomposition of Twitter disgust index 161](#_Toc502566532)

[Figure 5. 22 Wavelet Decomposition of Twitter fear index 161](#_Toc502566533)

[Figure 5. 23 Wavelet Decomposition of Twitter Joy index 162](#_Toc502566534)

[Figure 5. 24 Wavelet Decomposition of Twitter sadness index 162](#_Toc502566535)

[Figure 5. 25 Wavelet Decomposition of Twitter surprise index 163](#_Toc502566536)

[Figure 5. 26 Wavelet Decomposition of Twitter trust index 163](#_Toc502566537)

[Figure 5. 27 Wavelet nonlinear Twitter Emotion FTSE model structure 164](#_Toc502566538)

[Figure 5. 28 Wavelet nonlinear Twitter Sentiment FTSE model structure 165](#_Toc502566539)

[Figure 5. 29 Simulation results of daily FTSE & Twitter sentiment model 166](#_Toc502566540)

[Figure 5. 30 Simulation results of FTSE & Twitter sentiment model validation 167](#_Toc502566541)

[Figure 5. 31 Simulation results of daily FTSE & Twitter emotion model 168](#_Toc502566542)

[Figure 5. 32 Simulation results of FTSE & Twitter sentiment model validation 168](#_Toc502566543)

[Figure 5. 33 Simulation results of daily FTSE Twitter sentiment & emotion model 169](#_Toc502566544)

[Figure 5. 34 Simulation results of daily FTSE Twitter sentiment & emotion model validation 170](#_Toc502566545)

# Introduction

## Background

For modern technologies, modelling and forecasting of non-liner and non-stationary processes in different research areas is an essential method to improve industry management efficiency. The main feature of system identification is to construct a model to connect system inputs and outputs to reveal the relationship between these variables (Wei and Billings, 2004). Numerous systems can be approximately represented by simple linear or nonlinear system models, and non-stationary system modelling and analysis. For example, there are many real-world processes that are usually severely nonlinear and time varying (Billings, 2013). Wavelet is a mathematical function that describe a signal or time series in either time domain or frequency domain. This have made Wavelet theory can be widely applied in various areas, including signal processing and data modelling. Wavelet is defined as a mathematical model that is used to pre-process signals in nonlinear fields (Akrami et al., 2014). Wavelet-based models can be applied to reveal and characterize the inherent dynamics of non-linear and non-stationary processes. For instance, a wavelet-based model has been applied to forecast the monthly rainfall data in India (Masheswaran and Khosa, 2014); Kuo, Gan and Yu used a wavelet model to predict air temperature in Taiwan; Liu, Niu, Wang and Fan (2014), used wavelet transform and support vector machines to model the wind speed; Alquist, Kilian and Vigfusson (2011) used a wavelet-based model to forecast oil prices; a hybrid wavelet method has been used to model stock market process (Hsieh, Hsiao and Yeh, 2010); Wei, Billings and Balikhin (2004) applied wavelet models to measure the disturbance of magnetic storms; Wei and Billings (2006) also used wavelet models to predict water level; the Electroencephalography (EEG) signal can also be modelled by wavelet models (Wei et.al., 2010) (Li et al., 2012).

There is consensus that stock market prices have unexpected fluctuations both in the short and long term. Therefore, a reliable prediction method of the stock market could help investors to obtain profit when buying and selling. However, simulating a stock market is a challenge because such a financial time series is a complex process and its performance is influenced by numerous factors, such as political events, current and future economic conditions and the investors’ sentiments (Hsieh, Hsiao and Yeh, 2011). Existing models and forecasting methods about stock markets have limitations. More specifically, commonly used models are not able to deal with sharp changes or jumps in stock market systems. Therefore, more effective methods need to be developed. One method applicable for dealing with such severely non-linear processes is to use wavelet based models. By decomposing the system input variables in numbers of new time series at different levels (that is the approximation time series and the detailed time series), the complex system can be represented by a Wavelet Multi Input Single Output (WMISO) model. Generally, for linear WMISO model identification, the least squares method is an effective way to estimate the model parameters; for non-linear WMSIO model identification, the orthogonal least squares (OLS) algorithm and error reduction ration (ERR) test provide a good solution (Billings, 2013) (Wei and Billings, 2004) (Billings and Wei, 2005).

Traditional stock market analysis methods have usually applied regression methods to model stock market price volatility. However, such models have a technical flaw since the stock market change is influenced by political and economic factors together with potentially irrational behaviour of investors, which will make the model and predict results inaccurate. Behavioural economics considers that when modelling stock market volatility change, psychology and behaviour cannot be ignored. The study of investors’ behaviour has been impossible in the past, however, with the advent of big data, utilising the massive data from the Internet, to help a study stock market model, can be achieved. As the majority of Internet data is in text form, sentiment analysis algorithms will be used in this project.

## Motivation

Although there are many studies have shown that microblogging such as Twitter can provide numerous data for sentiment analysis (Pak and Paroubek, 2010) (Go, Bhayani and Huang, 2009) (Agarwal et al., 2011), the data extraction and collection is difficult, expensive and hysteresis. Hence, a methodology or platform that could help us to extract required tweets is necessary. Furthermore, tweets always contain different kinds of information, so tradition sentiment analysis methods cannot provide a good classification result for Twitter data. According to this, a methodology that can tidy and analysis the sentiment or emotion that is contained in Twitter data is very useful. In addition, Twitter as a platform for public’s information exchange, Twitter data include political and economic issue can be used to study the popularity variety in election (Wang et al., 2012) and economic models (Bollen, Mao and Zeng, 2011). System identification methods include linear and nonlinear model may not able to predict the political models and economic models. Therefore, novel techniques and methodologies should be developed to deal with Twitter sentiment analysis and complex non-linear non-stationary economic systems modeling. In general, Twitter data extraction techniques, sentiment analysis (machine learning and lexicon based methods) and complex system modeling algorithms are needed to be proposed and developed to handle economic and political systems.

This research is going to extract tweets from Twitter and implement a Lexicon-based method and Machine learning method to distinguish Twitter sentiments. The sentiment index will help us to study the political and economic systems. The Twitter data in this project has three categories of topics: the US presidential election (Donald Trump and Hillary Clinton); the UK referendum 2016 and the FTSE 100 closing price. The specific research problems are shown as follows:

* With the development of Information technologies, the Internet data has experienced an explosive growth. Although several online text data about economic and political are available, current data collection methods are proven to be inefficient and expensive. This thesis focuses on developing a novel method that is able to extract economic and political text information from social network service.
* Stock market price time series are severely nonlinear and include several significant uncertainties. Therefore, traditional nonlinear model or statistical analysis cannot capture the nonlinearity and the uncertainty of stock market system. The research focus on explore the applications of wavelet nonlinear methods in UK stock market system.
* This project plan to develop a novel algorithm that use online social network information to model and predict the UK stock market variances, Brexit 2016 and the US presidential election 2016. In this process, we will use state-of-the-art methodologies in signal processing, data mining, system identification and computational intelligence.
* Political events, such as Brexit 2016 and the US presidential election 2016, influence the world in different aspects. The public sentiment variety is of great significance in predicting the outcome of the referendum and election. As such, we focus on: mining daily Twitter sentiment variety of the two US presidential candidates Hillary Clinton and Donald Trump to predict the election results; mining the geography Twitter sentiment about Brexit to model the referendum.
* Another research problem is to compare the performance of different sentiment analysis algorithms (machine learnings and lexicon based). These algorithms will be improved and combined to develop a novel method that is suitable for analysing online text data at sentiment level.

## Overview

This thesis is mainly organized into six chapters: The research background and problem statement is contained in Chapter 1; a detailed review of the related theoretical research and methodology applications are discussed in Chapter 2; the process of exploring the applications of data mining and lexicon based method in sentiment analysis are included in Chapter 3; Chapter 4 has studied an advanced machine learning methods for sentiment analysis and complex network analysis for data visualization; in Chapter 5, sentiment data that we acquired from Chapter 3 and novel wavelet models are implemented to model and forecast stock market price and lastly, Chapter 6 present a detailed conclusion of this thesis and also provide the future research direction. The detailed thesis composition is shown below:

* Chapter 2

Chapter 2 mainly discusses the theories and applications related to this research, it gives us an in-depth literature review about three main problems of this thesis. 1. In this big data age, why and how the twitter data is able to influence the human life in political, economic and other aspects. 2. How to extract Twitter data and how to conduct a sentiment analysis of Twitter? 3. How to model non-linear, un-stationary complex system (such as Stock Market price) using system identification methods. Chapter 2 emphasis the Wavelet linear model and Wavelet nonlinear model, along with a review of the sentiment analysis methods for Twitter data: machine learning and NRC Lexicon based method. This chapter also discussed the applications of sentiment analysis and system identification methods in political and economic issues.

* Chapter 3

Chapter 3 deals with Twitter Mining problems and Opinion Mining (Sentiment Analysis) with R programming language. Twitter API is developed and implemented in R to mining Twitter data. NRC lexicon is used to classify the Twitter text data. In this chapter, three case study: Twitter US presidential election data, Twitter Brexit 2016 data and Twitter FTSE 100 close price data are made to show the Twitter opinion change.

* Chapter 4

Chapter 4 have proposed a novel feature selection methods for KNN and Naïve Bayes classifier. The Twitter data about Donald Trump is used to train and test the classification performance. Furthermore, complex network theory has been applied for two case study FTSE100 and R21-15 for the data visualization.

* Chapter 5

Chapter 5 developed linear Wavelet and nonlinear Wavelet models for non-linear and un-stationary system FTSE model. The main objective is to 1. explore and analysis if wavelet models can improve the predictive power of the FTSE system. 2. In this process, the relationship of popular world stock market index, crude oil price, Twitter sentiment index and stock market price are discussed and evaluated. More specifically, in the first case given world stock market indexes is implemented to train wavelet linear/nonlinear models and to predict the SSE composite index Changes. In the second case study, crude oil price is used as an input time series to train and test the FTSE price. In the third case study, Twitter sentiment indexes are used to model and predict the FTSE100 system.

* Chapter 6

Chapter 6 provide a detailed summary of this thesis, together with future research direction of this subject.

## Contributions

This project aims to study data mining, sentiment analysis and system identification approaches and applications. A novel algorithm that implementing Twitter data to model non-linear and non-stationary system is developed and this algorithm can be applied in either economic system or political system. The main contributions of this project are shown below:

* Chapter 3

Because the lack of research Twitter data and current data collection methods are inefficient and expensive, I developed my own Twitter API based on R languages. This program is able to help us to extract tweets based on keywords and tweets can be collected by geography location and post time. Tweets related to Brexit, US presidential election and FTSE are extracted. A total of 23332 tweets about UK Brexit referendum are collected; over 3 million tweets about US presidential election are collected; around 90000 tweets are extracted. These tweets are collected day by day. The value of these data is not only reflected in this study, but also important in other research area. Considering about the extracted tweets include unrelated information that will affect classification results. I also have proposed a novel text preprocess method for Twitter data. The preprocess method is able to distinguish tweets language, remove interference information and tidy tweets in order to deduce the bias for sentiment analysis. What’s more, in this chapter, I have explored the application of the NRC sentiment lexicon on Twitter. Daily Twitter sentiment/emotion variety about US presidential election, daily Twitter sentiment/emotion time series about FTSE100, geography Twitter sentiment of UK Brexit referendum are obtained. The Twitter US presidential election model can comprehensively reflect the public sentiment/emotion data towards these two presidential candidates. These time series data are significant for the future modeling and forecasting task.

* Chapter 4

A novel feature selection methods is proposed for Twitter opinion mining. Traditional feature selection methods such as Document Frequency (DF), Information Gain (IG) and Mutual Information (MI) have been widely applied text mining. I have proposed a new feature selection method which is applying the emotion features that acquired from NRC lexicon. These features will be applied to machine learning methods such as KNN and Naïve Bayes methods to classify the Tweets polarity (Positive or Negative). Results will be compared and evaluated to the traditional text feature selection methods in Twitter. Our experiment results show our hybrid model NRC lexicon and Machine Learning classifier have improve the performance of the classification results in Twitter.

* Chapter 5

Wavelet based NARX model has been firstly introduced to Stock Market price modeling. I implemented this method for three case study: 1. Europe stock market for word stock market prices for FTSE100 composite index system, 2. Weekly crude oil price & FTSE100 price system and daily crude oil price & FTSE100 system, 3. Twitter sentiment & FTSE100 system. Significant regressors terms that is able to describe the stock market changes are identified. The results show that Twitter sentiment/emotion index of FTSE 100 provide good validation result of the FTSE 100 daily close price.

# Literature Review

## Introduction

The aim of this research is to apply online information and datasets such as Twitter in addition to mathematical methods being used to model and forecast the UK stock market variation (FTSE100), important political events (the 2016 US presidential election, and the UK 2016 Brexit referendum). Considering the aim of this research, there are a few problems which should be explained and proved:

* How to acquire/extract online datasets from the Internet? Such as Twitter
* How to tidy and mine useful data from these information-rich text datasets?
* How can these online big datasets be classified and analysed?
* How to apply appropriate mathematical methods, nonlinear models to model a complex system, such as a stock market system?
* How to implement the online information to help to improve the predictive power of the stock market system?
* How to model the outbreak and spread of social behaviour and political events?
* How to predict the influence of some social and political problems?

According to Mao, Counts and Bollen (2011), Internet datasets include those from Twitter, news and search engine data. Recent research has demonstrated that search engine query data has been used to detect influenza epidemics (Ginsberg and H. Mohebbi, 2009). Furthermore, based on behavioural economics and Efficient Market Hypothesis, Twitter sentiment has been used to predict the US stock market (DJIA), gold price and other financial indexes (Bollen, Mao and Zeng, 2010). Internet information can not only be used to predict the financial market index and infectious diseases, but can also be applied to analyse social problems. For example, related research has shown that “during 2010 and 2011 Australian Floods, social network analysis of tweets” has successfully developed an online community (Bird, ling and Haynes, 2012). Furthermore, the importance and disseminating of this community has been identified (Cheong, 2011). In this chapter, the research mainly focuses on modelling mathematical, web data mining, sentiment analysis. All these will be represented and how this is related to the present research will be discussed.

## Modelling and Forecasting Methods

### Introduction

The aim of this project is to model nonlinear and non-stationary systems using advanced system identification methods. In order to successfully model and predict complex nonlinear processes, two important factors need to be considered. Firstly, appropriate models should be applied such as linear models, nonlinear models, neural network models, statistical models and hybrid models. Second, the inherent properties of the system should also be considered and analysed. There is much research related to modelling and forecasting stock market indices. Mining stock markets is a challenge. Analytical indices that have been proposed contained “price multiples, macro variables, corporate actions and measures risk” (Ferreira and Santa-Clara, 2011). In this literature review, different modelling algorithms for stock markets will be discussed and compared.

### Linear Models

Models are a significant part in system design and analysis. “System identification is a technique that can be used to infer and construct system models from experiment data” (Billings, 2013). In order to model a system, different types of linear model such as autoregressive (AR), autoregressive with exogenous input (ARX), autoregressive moving average (ARMA), autoregressive moving average with exogenous input (ARMAX) models can be implemented. However, most of the real-world systems are nonlinear or even severely nonlinear; linear models are not able to capture important inherent dynamics for example the “rich dynamic behaviour of limit cycles, bifurcations” (Rahrooh and Shepard, 2009). Ferreira and Clara (2011) used a regression method to model stock market returns and found that linear regression cannot provide an accurate prediction result. Campbell and Thompson (2005) also state that linear regression is not reliable in a stock market return model because “estimated parameters are not stable over time”.

### Nonlinear Models

Nonlinear systems are defined as systems that are not linear, which means that a system does not satisfy the superposition principle (Billings, 2013). Numerous applications of nonlinear models in real-world problems have proven that nonlinear models can improve the prediction accuracy compared with linear ones. According to Chen and Billings (1989) many nonlinear systems can be represented by the NARMAX model. It has been shown that numerous real-world systems can be modelled by NARMAX (Chiras et.al, 2001) (Fung et.al, 2003) (Jain and Kumar, 2007) (Deng and Tan, 2009). The mathematical representation of the NARMAX model is shown in the equation below (Billings, 2013).

(2.1)

In the equation above, is a nonlinear function, is the system output, is the system input and is the noise term. is the maximum lag of output, is the maximum lag of input and is the maximum lag of noise term. The model output is defined by its past values, noise and exogenous input.

According to Billings (2013), nonlinear systems include mildly nonlinear systems and severely nonlinear systems. Many engineering systems belong to mildly nonlinear systems that are stable and can be modelled by NARX or NARMAX models (Billings, 2013). With widely used of systems identification technologies, Increasingly, real-world systems have been considered, such as stock market systems, oil price systems meteorological and hydrological systems. These systems are nonlinear, complex and non-stationary, and for such severely nonlinear systems, polynomial NARX and NARMAX models may not be enough to provide satisfactory prediction results.

### Granger Causality Test

The Granger Causality test (GC) describes whether a time series is significant in forecasting another time series using a hypothesis test. GC is a noted algorithm for causality tests. Additionally, the algorithm is usually applied in a vector auto-regression (VAR) context. The algorithm can be used for causality prediction. More specifically, “Granger Causality test is a statistical hypothesis test to determine whether a time series is useful in forecasting another time series by attempting to reject the null hypothesis that does not help predict” (Mao, Counts and Bollen, 2011). Venezia, Nashikkar and Shapira (2011) also found that numerous results have shown that Granger Causality can provide some understanding of predictability.

When discussing the Twitter sentiment prediction for a financial market, previous research in this area has utilized news and surveys to acquire the sentiments of investors (Mao, Counts and Bollen, 2011). Using large-scale online data, such as Google surveys, Twitter and Facebook, to acquire public sentiment has become a trend in research studies. For example, Bollen, Mao and Pepe (2011) applied Granger Causality test to explore the cross correlation of Twitter Opinion and the Dow Jones Industrial Average (DJIA). In order to test if Twitter Opinion time series data is able to predict changes in the DJIA time series index, researchers built two linear models shown below (Bollen, Mao and Pepe, 2010):

(2.2)

As shown in the equations above, is the DJIA index and is the Twitter Opinion data. According to Bollen, Mao and Pepe (2011), the first model only uses the delayed DJIA index as the predictor while the second model uses both n delayed values and Twitter Opinion time series for prediction. Based on the results from Bollen, Mao and Pepe (2011), the null hypothesis that Twitter Opinion cannot predict the DJIA index should be rejected. Furthermore, the Granger Causality test shows that calm sentiment is Granger-Causative of the DJIA index (Bollen, Mao and Zeng, 2011). Mao, Counts and Bollen (2011) concluded that “the predictive power of Twitter’s two sentiment indicators outperformed survey sentiment as well as news media analysis.”

Granger Causality is a popular method to reveal causality influence of two time series based on linear regression models and is widely applied in economics (Hu and Liang, 2012). There is some debate on whether Granger Causality is appropriate stock market systems since stock market prices will be influenced by numerous factors (Eichler, 2012). More specifically, if the input time series and output are also influenced by another variable with different lags, the Granger Causality may fail to reject an alternative hypothesis. Furthermore, normal Granger Causality will only reflect the linear feature of the time series and stock market systems are known to be complex nonlinear systems. Lastly, the stock market is a non-stationary system, whilst Granger Causality is defined as the analysed time series being covariance stationary (Eichler, 2012). In conclusion, a normal Granger Causality test is not suitable for stock market prediction.

### Artificial Neural Network

An Artificial Neural Network (ANN) is a statistical model that can be used to model complex systems with various numbers of unknown inputs. The architecture of an ANN consists of nodes and their connectivity. Generally, ANN architecture is described by the network inputs, network outputs, the number of nodes, organization and the interconnections (Kantardzic, 2011). Kantardzic (2011) also states that the ANN architecture can be classified into feedforward and recurrent. The operation of the feedforward ANN is unanimous, which means there are no feedbacks or loops. Feedforward ANN always consists of three layers, namely input layer, hidden layer and output layer and all these are completely connected to build a hierarchical network. More specifically, the input variables are imported simultaneously into the input layer. Then, after processing, the output from the input layer is imported simultaneously to the second layer, known as the hidden layer. Then the output of the hidden layer is the input making up the output layer and these are the predictions for the system (Enke and Thaworn, 2005). If the feedback or circular path appears, then the ANN is recurrent. Examples of ANN architectures are given in the figures below (Kantardzic, 2011).

Inputs

Hidden Layer 1

Hidden Layer 2

Hidden Layer 3

Output Layer

Figure 2. 1 Feedback ANN architecture

Delay

Inputs

Outputs

Figure 2. 2 Recurrent ANN architecture

For the current research and applications, 90% of the ANN models are based on the multilayer feedforward architecture. Implementing multilayer rather than a single layer has been done because ANN with a single layer is convenient for modelling simple linear classification problems. In real-world problems, the systems are usually complex, nonlinear and un-stationary, hence multilayer ANN are better than single layer ANN.

A stock market can be regarded as a non-linear, dynamic complicated system (Tan, Quek and Ng, 2005). Furthermore, stock markets’ changes are affected by many macro-economic factors such as worldwide political and economic issues, investor sentiment, stock market movements, commodity prices and economic conditions. Several studies using nonlinear models have proven that there is a connection between a neural network and stock market index. Enke and Thawornwong (2005) suggested that because many modelling techniques are linear models, a nonlinear model analysis about a stock market index should be considered. Enke and Thawornwong (2005) also state the two advantages of neural networks:

* Since neural networks can learn the inherent relationship of the variable independently, the method does not have a pre-specification process.
* Neural networks provide numerous and flexible “architecture types, learning algorithms and validation procedure.”

Zhang and Wu’s (2009) research applied an “improved bacterial chemotaxis optimization (IBCO)” algorithm integrated into the “back propagation artificial neural network” to develop a prediction model. A back propagation neural network (BP) is a supervised learning model. The basic principle of BP is using “the steepest gradient descent method” to achieve estimated approximation (Zhang and Wu, 2009). Similar to the typical artificial neural network, there are three layers in a BP network, which are input layer, hidden layer and output layer. There exists a link between each two nodes (Zhang and Wu, 2009).

The bacterial chemotaxis optimization (BCO) is an algorithm proposed by an “analogy to the way bacteria react to chemo-attractants in concentration gradients (Zhang and Wu, 2009).” When processing a BCO algorithm, firstly, the velocity of bacterium should be computed. Secondly, it is necessary to compute the trajectory using the exponential probability density function. Then, the new direction, referring to the previous trajectory, should be determined. Lastly, the new position can be easily acquired. The result of Zhang and Wu’s research (2009) shows that the stock market index can be predicted using a BP neural network. Performance of the IBCO model can provide better prediction accuracy than before. There is another method using a hybrid approach, based on ANN theory. More specifically, according to Kim and Shin (2007), using adaptive time delay neural networks and the time delay neural networks with the genetic algorithms in predicting the stock market is more effective. Based on Cao et. al (2005), a comparison between linear models and neural network models in SSE composite index is implemented. The results indicate that the predictive power of a neural network is better than for linear models and a neural network is an effective method in modelling stock markets (Cao, et. al, 2005). However, neural networks are not perfect in modelling severely nonlinear systems.

### Wavelet Pre-process for Nonlinear System Identifications

Wavelet is a mathematical function that describes a time series in time and frequency domain. Wavelets are a particularly useful method for localized approximation of functions at different frequencies. More specifically, wavelet makes it possible to use long time intervals to show high scale information and short time intervals to show low scale information. Different from Fourier transform, wavelet can be localized either in a time domain or frequency domain. Compared with other resolution methods, wavelet decomposition provides a practical and flexible method to approximate severely nonlinear signals.

A novel approach that implemented wavelet multi-resolution decompositions and system identification method to model severely nonlinear systems was first proposed by Wei and Billings (2002). The main feature of the wavelet method is the “stepwise algorithm used to derive the sparse representation of the unknown nonlinear system with minimum computational cost” (Billings, 2013). Many properties have made wavelet models an ideal method for severely nonlinear system identification. However, the wavelet models’ performance in real-world severely nonlinear systems should be considered.

Wei, Billings and Balikhin (2004) first introduced the wavelet identification models to predict the Dst index; “Dst index is used to measure the disturbance of the geomagnetic field in the magnetic storm”. A previous study predicting the Dst index used ARMA and NARMAX models, however, in order to obtain a better predictive power, the wavelet-based nonlinear model is introduced (Wei, Billings and Balikhin, 2004). The results show that the wavelet nonlinear model has a good predictive power, which is better than other approximation schemes. Therefore, wavelets are proven to be an effective tool in nonlinear system identification (Wei, Billings and Balikhin, 2004). Wei and Billings (2007) also state that the wavelet identification method outperforms iterative models for nonlinear time series. Wavelet multi-resolution can be used in neural network models. In Adamowski and Sun (2010), a wavelet artificial neural network (WA-ANN) was applied to model the flow of “non-perennial rivers in semi-arid watersheds”. The result showed that, for 1 day ahead and 3 days’ ahead prediction, the WA-ANN model outperformed regular neural network models. Wavelet models are more reliable because wavelet transform provides a more accurate resolution of original signals and captures more effective information at decomposition levels (Adamowski and Sun, 2010). A wavelet neural network model has been used to forecast the stock market (Hsieh, Hsiao and Yeh, 2011). There is another application of a wavelet-based model for predicting rainfall data in India (Maheswaran and Khosa, 2014), where the wavelet volterra model, wavelet linear regression model and wavelet volterra model outperformed other models. The reason that wavelet linear regression and neural networks can perform well for a hydrological system is because a linear model is not able to capture the nonlinear feature of systems and neural networks cannot pick up the nonlinearity of the system (Maheswaran and Khosa, 2014). However, whether these models are suitable for stock market still needs to be explored.

1. Reasons for Using Wavelet

The most significant objective of a nonlinear system identification is to obtain an appropriate model based on input and output variables. This procedure can be described as applying polynomial functions, kernel functions and other basic functions with global or local characteristics to construct a nonlinear model. According to Wei, Billings and Balikhin (2004), most types of functions can only be used to approximate certain severe nonlinear behaviour effectively. Furthermore, in some cases, the nonlinearity of the dynamical system cannot be represented at all “by a given class of functions because of the lack of good approximation properties”. It is generally recognized that, the basic functions used for the purpose of approximation should provide some flexibility “in adapting to the complexity of the model structure so that the model can match, as closely as possible, the underlying nonlinearity of dynamic systems” (Wei, Billings and Balikhin, 2004).

When the wavelet analysis was first introduced by Morlet and Grossmann in 1984, it was purposefully created to have the capability that incorporates the global basic function feature and local basic function feature that can be applied in signal processing. Wavelet outperforms Fourier transform and is suitable for arbitrary signals, such as severely nonlinear signals. Fourier transform only explains the frequency domain information and the time information is lost, hence, it is impossible to know when a specific change of signal takes place. Compared with Fourier transform, Wavelet transform has the ability of resolution and localization, which could transform and analyse signals both in frequency and time domain; this could overcome the defect of Fourier transform. Wavelet analysis applies a prototype function, called mother wavelet, which is used to decompose a signal into different scales.

1. Selection of Wavelet

In wavelet analysis, there are different kinds of wavelet functions used in practice. The results of applying different wavelet functions for data analysis may be different. Implementing the right and proper wavelet is a crucial step for wavelet transform (Megahed, Moussa, Elrefaie and Marghary, 2008). There are no general or standard methods for wavelet selection for a specific area.

Normally, selection of an appropriate wavelet function requires the understanding of wavelet properties, such as the wavelet support region, wavelet vanishing moments, similarity and symmetry (Ngui, Leong and Hee, 2013). According to previous research, it is necessary to consider the properties of a mother wavelet in the procedure of mother wavelet selection. Generally, more than one mother wavelet meets the requirements of a signal’s decomposition process. Therefore, “the similarity between the processed signal and mother wavelet should be considered in selecting a mother wavelet” (Ngui, Leong and Hee, 2013). Given this, the properties of mother wavelet and the similarity between a signal and mother wavelet are two important factors in choosing a mother wavelet.

More specifically, Ngui (2013) states that although there is no general and standard method in mother wavelet selection, the mother wavelet selection procedure can be based on qualitative approaches and quantitative approaches. Considering the regularity and vanishing moment of the mother wavelet, Mojsilovic et al. (2000) applied a biorthogonal wavelet for texture characterization. Fu et al. (2003) implemented the mother wavelet biorthogonal 6.8 to decompose the surface profiles by the symmetry properties. Compact support property and vanishing moment property have been used to select the proper wavelet in power system transients (Safavian, Kinsner and Turanli, 2005). Wang et al. (2004) conclude that the properties of vanishing moment, orthogonality and compact support are important in EGM signal decomposition. For an image processing area, the main properties of the mother wavelet are orthogonality, compact support, symmetry and filter order (Ahuja et al., 2005) (Adamo et.al., 2013). As discussed above, the similarity of the mother wavelet and the signal also contribute to the wavelet selection process. More specifically, the visual shape matching is widely implemented to pick the optimal mother wavelet (Ngui et al., 2013). Tang et al. (2010) find that the morlet wavelet is quite similar to the mechanical impulse signal, and the wavelet is implemented to denoise vibration signals. Furthermore, the db2 wavelet is found to be very similar with EMG signals. Flanders (2002) use db2 to measure the timing frequency of EMG signals. Ahadi and Sharif (2010) investigate whether the gauss mother wavelet is similar to the acoustic emission leakage signal.

There are more accurate methods that can help to measure the similarity between the mother wavelet and signal. These are quantitative approaches for mother wavelet selection (Ngui, 2013). For this current research, some special algorithms have been used in different areas of wavelet feature extraction. According to Zhang’s (2005) research, information extraction criterion and distribution error criterion are used to choose the proper mother wavelet for image denoising. The MinMax information criterion is implemented to acquire the most suitable wavelet in bearing faults’ detection (Yan, 2007) (Kankar et al., 2011). Considering biomedical research, Phinyomark et al. (2009) implement the wavelet coefficient’s mean square error (MSE) for the EMG signal decomposition. Phinyomark et al., (2009) also states that there are two most widely used algorithms can help to justify the similarity between mother wavelet and signal, namely: Minimum Description Length (MDL), Partial Discharge (PD).

* Minimum Description Length

“MDL is an algorithm that suggests that the best model among the given collection of models is the shortest description of the data and the model itself” (Ngui, 2013). MDL has been applied into signal compression, noise suppression and power disturbance data (Satio, 2004) (Effrina et al., 2001).

* Maximum cross correlation coefficient criterion

The maximum cross correlation coefficient criterion has been successfully applied in Partial Discharge signal extraction and ECG signal to find the optimal mother wavelet (Li, 2009).

More detailed study of the selection of wavelet for the stock market process will be implemented in future research. For the present study, the DB2 wavelet function is applied for wavelet decomposition.

1. Implementation of Wavelet Models

Wavelet theory has been widely used in signal processing and data modelling. When the wavelet models were first proposed by Wei and Billings (2002) to process non-linear and non-stationary systems, they were used to measure the disturbance of a magnetic storm (Wei, Billings & Balikhin, 2004); to predict water level (Wei & Billings, 2006); to model the Electroencephalography (EEG) signal (Wei, et.al., 2010); to forecast the oil price (Alquist, Kilian and Vigfusson, 2011); to predict rainfall data in India (Masheswaran & Khose, 2014); to predict air temperature in Taiwan (Kuo, Gan & Yu, 2010).

## The Influence of Twitter Sentiment

### Background

Nowadays, people are more dependent on the Internet than ever before and the Internet has profoundly influenced our daily life. For example, people need the Internet to contact their friends, to shop for what is needed in their daily lives, for browsing web pages they are interested in or even post their feeling and images on Twitter or other public social media. More specifically, modern people live in a world where human behaviour and activities will leave digital traces and these traces will affect people’s daily life (Bordino et.al, 2012). A digital trace of people’s daily life can include online records, online comments, search engine data and web browsing history. Online records (online shopping record, downloaded records and bill records), search engine data (what people search for on Google/Baidu) and web browsing history are related to personal privacy and are not considered in this research. Online comments can include Facebook comments, tweets on Twitter and other comments on the web. People on the Internet can be easily affected by other users’ articles and reviews. Text based datasets can be easily found on Twitter, Facebook and YouTube. Twitter is a website and social media platform with a large number of text datasets that can be used for “opinion mining and sentiment analysis” (Pak and Paroubek, 2010). Some research has been carried out and the text dataset/comments are proven to be able to help with analysis and prediction studies. In other words, numerous studies focus on using the comments and online text datasets/comments that people post on the Internet to model and predict specific information, such as applying sentiment and search query to predict the box office of movies (Mao et.al, 2011), online sentiment to predict financial markets (Mao et.al, 2011), Internet search queries to predict stock market volatility (Bordino et.al, 2012), search engine data to detect influenza epidemics (Ginsberg et.al, 2008). As such, the development of social media research presents “a great opportunity to understand the sentiment of the public via analysing its large-scale and opinion-rich data” (Hu, et.al, 2013). Hence, in this research, the main focus is comments on one of the most popular social media platforms, Twitter.

Web information contains different kinds of data forms, such as online records (online shopping records, downloaded records and bill records), online comments (Twitter, Facebook and YouTube), search engine data and web browsing history. As discussed before, online records, search engine data and web browsing history are related to personal privacy issues, so it is difficult to legally acquire and analyse these data. Furthermore, online records, search engine data and web browsing history information are always opinion-deficient, hardly mined and small-scale. These various forms of data cannot fully reflect people’s sentiment information and can be difficult to use for future modelling and forecasting. Twitter, as a popular and worldwide social media platform, provides information about different affairs, which can be acquired for different purposes and its large dataset is a key factor since it can be used to model a real-world system. For example, Twitter includes sentiment data about stock markets and other financial markets. In the next section, the importance of Twitter will be discussed in detail.

### What Makes Twitter Sentiment Significant

First the important issue of why Twitter contents are utilised in this present study. It has become a popular trend that large numbers of investors post their opinions, attitudes and comments about recent stock trends through their Twitter account. Additionally, popular newspapers also have Twitter accounts that are focused on stock markets. Investors and media moods about stock markets can be easily spread and influence others through Twitter and this online platform allows people to post their views about stock markets using no more than 280 characters. Due to behaviour economics, Twitter sentiments could reflect investors’ mood about stock markets and this may influence the markets.

There is another reason for using Twitter information. Twitter datasets are time-scale data. More specifically, each Twitter user has access to other users’ tweets with no limitation. Due to the time varying, different investors will post tweets on Twitter, which means that researchers can obtain the time-scale updated data to analyse the public mood and predict stock market variety and change. It could mean significant progress if Twitter contents can be tracked and extracted to obtain the real time public mood information about a stock market. This method can not only be used to model stock market systems, it could also be applied to political and economic models.

Tweets in Twitter contains much information that can be worth mining and analysing. The information has many internal links that can help researchers to model and forecast the economic and political changes. In this research, tweets related to the 2016 US presidential election (Hillary Clinton and Donald Trump), UK Brexit 2016 and the UK stock market (FTSE 100 index) are considered.

Tweets on Twitter are important in predicting stock market trends. Stock markets can be regarded as a system with an input of Twitter feeds. Based on this, Twitter mood analysis is becoming a trend to predicting the economic system (Bordino et.al, 2012). Some theories have also shown the importance of sentiment information. In behavioural economics theory, emotion is able to influence human beings’ behaviour and decision-making (Billen, Mao and Zeng, 2011). Prechter (2002) also states that social mood is an important factor that can influence financial decision-making. The traditional method for predicting the stock market is based on public mood data from survey and news. However, these sentiment data are not time updated. More specifically, surveys can only be acquired by people who take part in the activities and complete the survey papers, after that, these survey papers still need to be processed and analysed before the final results are clear. Similar to survey, news needs to be collected and analysed before it is applied.

Worldwide political and economic events always have a significant and profound impact on stock market systems. These might include, news like political issues such as the US presidential election, Brexit, the EU debt crisis and other worldwide political issues. With breaking news, investors’ sentiment can also influence the stock market. According to Bollen (2011), public sentiment played a significant role on market decision-making. Nofsinger (2005) stated that “behavioural finance has provided further proof that financial decisions are significant driven by emotion and mood.” As such, there is evidence to show that investors’ attitudes can profoundly influence the stock market trend.

To explain why Twitter datasets are used, first Twitter is an online platform where users can post tweets in no more than 280 characters. These tweets usually include their attitude towards a topic or something they are interested in. Based on the theory of behaviour economics, Twitter sentiment is able to reflect investors’ mood about stock markets and this will have profound influence on the markets. As discussed previously, large numbers of investors post tweets about stock markets via their Twitter accounts. Furthermore, popular newspapers also have Twitter accounts that are focused on stock markets’ changes. Investors and media sentiments about a stock market index could easily influence other Twitter users.

### Twitter Network Communication Analysis

With the dramatic development of the mobile terminal (MT) in recent years, researchers from different areas have tried to study Twitter from different perspectives. This part will study the communication and interactive behaviour in Twitter users and how Twitter information is transmitted.

There is one situation that always happens: some of the Twitter users may have more than one Twitter accounts. This means that a user may have different intentions; they may use these accounts to support their ideas and play different roles when communicating with other Twitter users in different social networks. This pattern of social network communication could generate numerous Internet links and datasets, which would be meaningless in this research. It is useful to understand the pattern of the Twitter data dissemination, and the Figure 2.3 below shows the simple pattern of Twitter data dissemination.

Figure 2. 3 Simple Twitter dissemination process

The most significant process of Twitter communication is for a tweet to be forwarded. Once a tweet has been posted by a user, the user’s friend can forward the tweets if they think it is interesting or they want to show it to their friends. The forwarded relationship is illustrated in Figure 2.3 above. Figure 2.3 shows that Tweet X is forwarded by Twitter users A, B and C. Next, the tweet has been forwarded by A, B and C’s friends D, E, F, G, H.

The influence by a Twitter user to other Twitter users is interactive. Breaking news also needs Twitter to make it become significant through Twitter posts, forwarding and discussion. There can be two reasons why some breaking news do not have a wide range of social impact: 1. There may be another news storying happen which is more significant than this news. 2. Few Twitter user are involved in commenting or forwarding. The number of forwarded times, the number of replies received and the number of people viewing are three quantitative criteria for tweets. According to Cha (2010), popular Twitter users (people with many followers or whose tweets are forwarded heavily) do not necessarily directly influence their followers and topic participants. Meanwhile, Romero’s research (2010) illustrates that the influence of a tweet is not only decided by the Twitter user’s prestige. More specifically, the relationship between the Twitter user’s prestige and the influential power of the user’s tweets is weaker than expected, as most Twitter users are not able to filter the tweets that are posted by their friends. Yang et al’s research (2010) research shows that three key factors that can help to study and build the Twitter spread model. They are: 1. Spread Speed; 2. Spread Scale; 3. Spread Range. These attributes of Twitter can help to construct the Twitter data spread model.

Behaviour economics considers that sentiment can influence individuals’ behaviour and decision-making. The relationship between social network sentiment and economics has been a key issue in recent years. Increasingly, tweets' discussion relating to economics problems have been posted. These opinion-rich tweets are spread through social networks and influence the public sentiment. According to behaviour economics, either optimistic and pessimistic sentiment tweets will indirectly influence the world market economy.

### Web Mining

“Web mining is the application of data mining techniques to discover patterns from the World Wide Web” (Cooley, Mobasher and Srivastava, 1997). Web mining is a combination of several research area such as Information Retrieval (IR) and Information Extraction (IE), but there are differences between them (Kosala and Blockeel, 2000). IR is a sample of Web content mining; intelligent IR is Web Mining. The purpose of IR is finding useful documentations in a text index and large data collection such as the Internet. Currently, the area of IR includes modelling, text classification, text clustering, user interface, data visualization and filtering. The main purpose of Web mining is web text classification and clustering. From one point of view, Web mining is a part of information extraction, however, not all IR tasks will use data mining technologies. IE is a process implementing the information/data, which is acquired from IR, to process data. More specifically, IE focuses on extracting the facts from documentation, IR focuses on extracting related documents. IE mainly deals with the knowledge, structure and expression of a document. The Internet contains a variety of forms and information and as there are many IE systems designed for specific websites this will lead to poor system scalability.

According to the three types of data (categorical data, sequential data and numerical data), Web mining can be divided into three categories: Web content mining, Web structure mining and Web usage mining, as shown in Figure 2.4 below:

Web Mining

Web Content Mining

Web Structure Mining

Web Usage Mining

Text Mining

Media Mining

Hyperlink Analysis

Webpage Structure Mining

Media Mining

Analysis Customization

Figure 2. 4 Web Mining Systematics

Web content mining is related but different from data mining and text mining, it requires data mining techniques and creative application of text mining techniques (Pol et al., 2008). According to figure 2.4, web content mining can be divided into text mining and media mining. The former is based on the data base and data mining technologies, such as inductive, classification and clustering, the object of web text mining can be either structured or unstructured (Liu, Hu and Cheng, 2005). The result of web text mining can be either the generalization of a specific text content, or the classification and clustering result of the entire text collection. Currently, Web text mining mainly focuses on summary, classification, clustering and relation analysis of the large collection text on the Web (Pol et al., 2008).

### How to Extract Tweets on Twitter

There are many methods to obtain the UK FTSE100 index but a useable and effective method to access tweets is not that easy. In this research, the UK stock market index FTSE100 has been acquired from Yahoo finance, while the Twitter contents were extracted from <http://twitter.com/>. It is difficult to extract useable and relevant sentiment tweets dataset from Twitter. Hence, Twitter content mining, Twitter sentiment analysis is the key factor for stock market prediction. Although Twitter has the search function which can help researchers to find the most relevant tweets, this function can only provide the access for the past three days’ data, which means that these sentiment data should be collected at least every 3 days. Therefore, an effective and practical method for mining and analysing tweets’ contents should be explored.

Real-time Tweets content data is necessary for the stock market system. Therefore, an automatic and practical way to undertake the Twitter mining process is important for this research. In this experiment, daily Twitter data is needed for analysis and these tweets should be well structured for the analysis. The structure should include contents, author name and post time. In order to make the process of tweets extraction easier, Google SpreadSheets, Webharvey and R program are implemented and compared.

Google SpreadSheets is an online spreadsheet program supported by Google Drive. Similar to Microsoft Excel, Google Sheets can perform many functions such as calculation and simple programming. Based on these features, the researcher used the *ImportXML* function in Google Sheets to execute the Twitter extraction process. There are three factors of a tweet that are significant for this research: time, author and content. However, there are some limitations of this method. Firstly, due to the Twitter search protocol on <http://twitter.com/>, only 100 tweets can be acquired from this method. Secondly, the post time of extracted tweets is not updated. The reason for this is not clear. Given the drawbacks of Google Spreadsheets for Twitter extraction, the tweets extracted using Google SpreadSheets are not reliable and cannot be directly applied.

Another method is using Webharvey which is software that can extract online data (URLs, text and images) from web pages automatically. Webharvey can also save the extracted data in different formats (<http://www.webharvey.com/>). It seems that Webharvey could meet the requirements of Twitter extraction. However, there are still some drawbacks with the Webharvey application. The main problem of Webharvey is the post time of each tweet. As the Twitter post time is not complete in the search page, it only has the time from when the tweet was posted until the present time. Another drawback of Webharvey is that the software can only show the first page of Twitter extraction results in miner data dialog. The third problem is that Webharvey cannot do the mining process automatically and researchers still need to undertake some procedures to finish it.

R language is also able to help researchers to extract Twitter messages. The extracted tweets can be stored in either Microsoft Excel or Word format. Furthermore, R language can mine the tweets’ information based on the location of the tweets. This type of information does not just reflect the location of Twitter information, the geographic information allows researchers to observe the variety of the Twitter public sentiment from a comprehensive perspective. The advantage of R language in Twitter information mining is not just about the comprehensiveness of the information extraction, R can also help researchers to deal with the extracted Twitter data with its own language processing packages. However, R language still has its defects in Twitter mining; R cannot perform daily extraction tasks automatically, researchers need to extract Twitter data day by day. Compared with Webharvey, R cannot obtain the author information from each tweet and this will influence the analysis of complex network systems.

In conclusion, considering the requirements of this project, neither Google Sheets nor Webharvey are ideal methods for Twitter extraction compared with R language. R language’s advantage is mainly in its high extraction efficiency namely more tweets’ collection and a powerful language processing system. Hence, the programming language R will be implemented in our future research.

1. Google Spread Sheet

Google Spread sheet is based on Google Drive and many functions can be imported and programmed. Accordingly, , the Twitter extraction process required *ImporXML* function. As discussed in the literature review, three factors of a tweet that are significant for this research: post time of tweet, author of tweet and tweet content. Based on this, a spread sheet was established as shown in Figure 2.5 below.

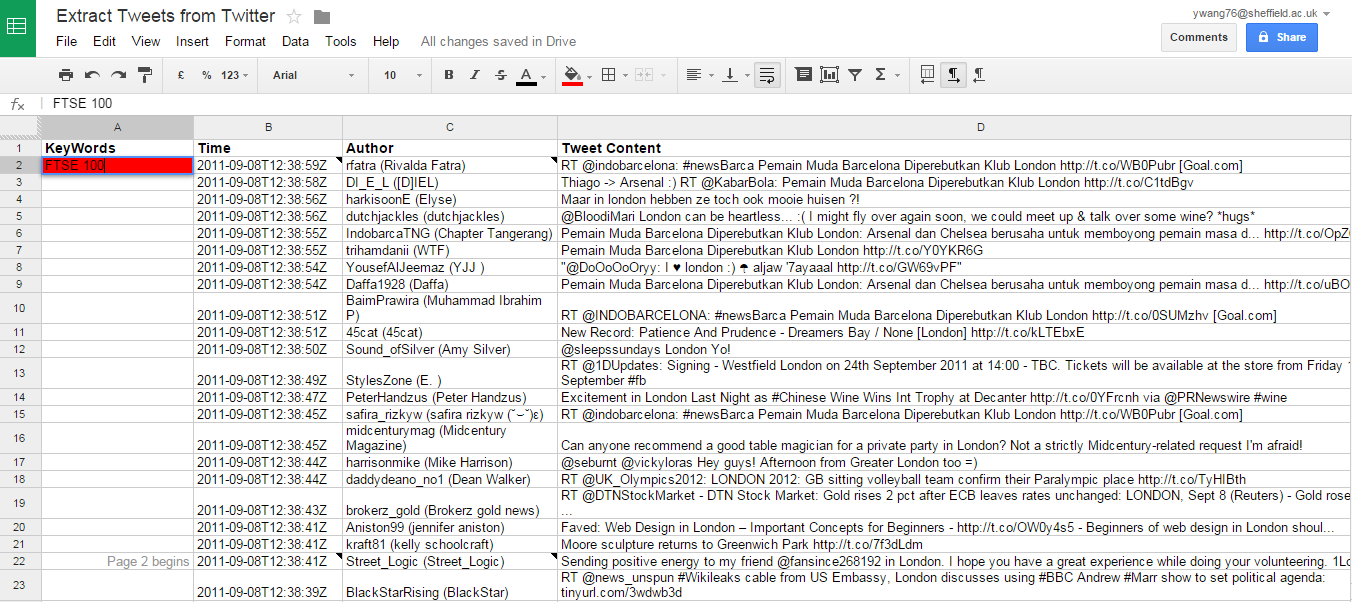


Figure 2. 5 Google Spread Sheet for Twitter Extraction

As shown in Figure 2.5, users can extract by simply typing the keywords in the red square, then the relative information will be extracted. It can be seen from Figure 2.5 that the post time, tweet author and content are acquired and can be saved in *.xlsx* form. However, there are some limitations to this method. Firstly, due to the Twitter search protocol on [*https://twitter.com/*](https://twitter.com/), only 100 tweets can be acquired from this method. Secondly, the post time of extracted tweets is not updated. The reason for this is unknown and the researcher will continue trying to solve it.

1. Webharvey

Considering about the limitations of Google Spreadsheets with Twitter extractor, the tweets extracted using Google Spreadsheet is not applicable and cannot be used for Twitter extraction. Another software called Webharvey can be used for Twitter content extraction. Webharvey is able to extract online data in different formats, such as storing the online dataset in Excel format. How to collect tweets from Twitter by Webharvey will now be shown.

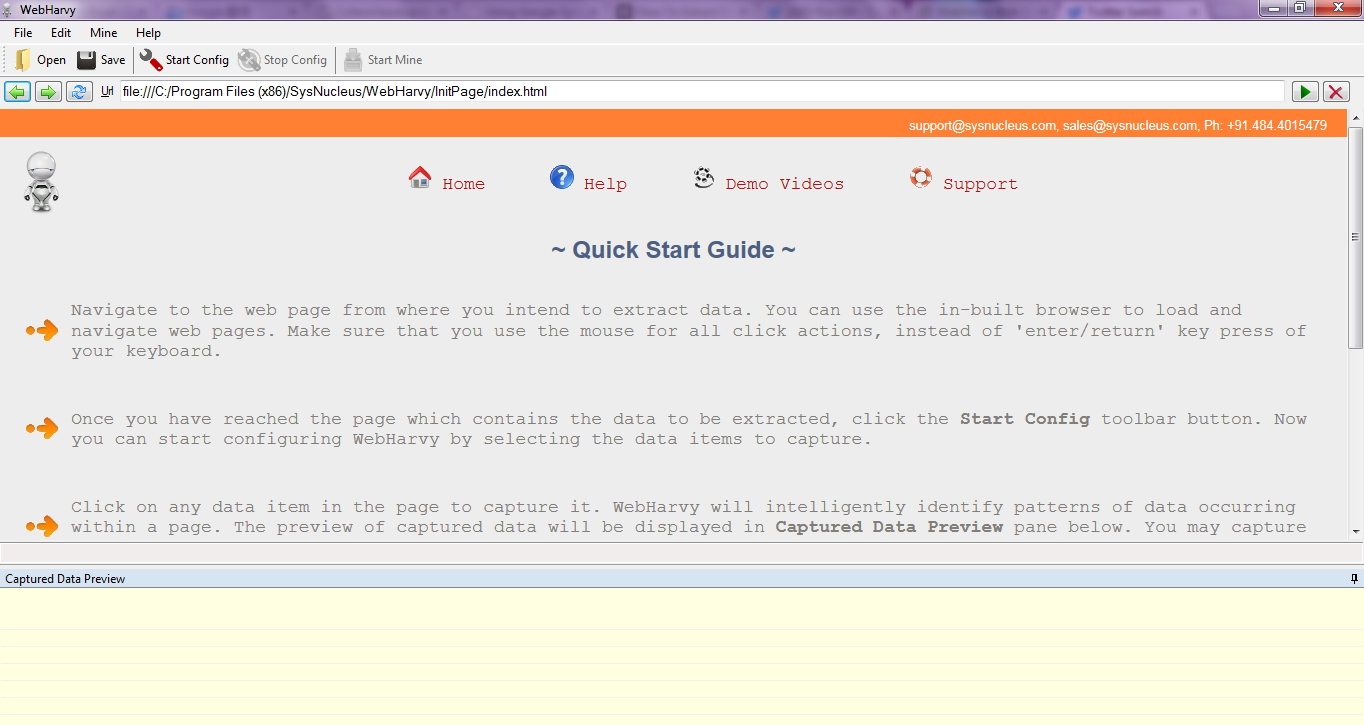


Figure 2. 6 Webharvey Operator Interface

Figure 2.6 above is the operator interface for Webharvey. The operating procedure of Webharvey is: Firstly, input <http://twitter.com/search-home>into the input bar at the top of the interface. The next step is to access the Twitter search page at the Webharvey interface and enter the keywords about stock market or a topic that is needed for the sentiment analysis in the search bar. After this, press the “Start Config” button. Researchers can acquire information about tweets’ content, tweets’ authors and post time. By repeating the process, researchers will obtain information about post time and author. The final step is to click on the “Stop Config” button and then the “Start Mine” button can be pressed. For example, using the keywords “FTSE100” the result is shown in Figure 2.7 below.

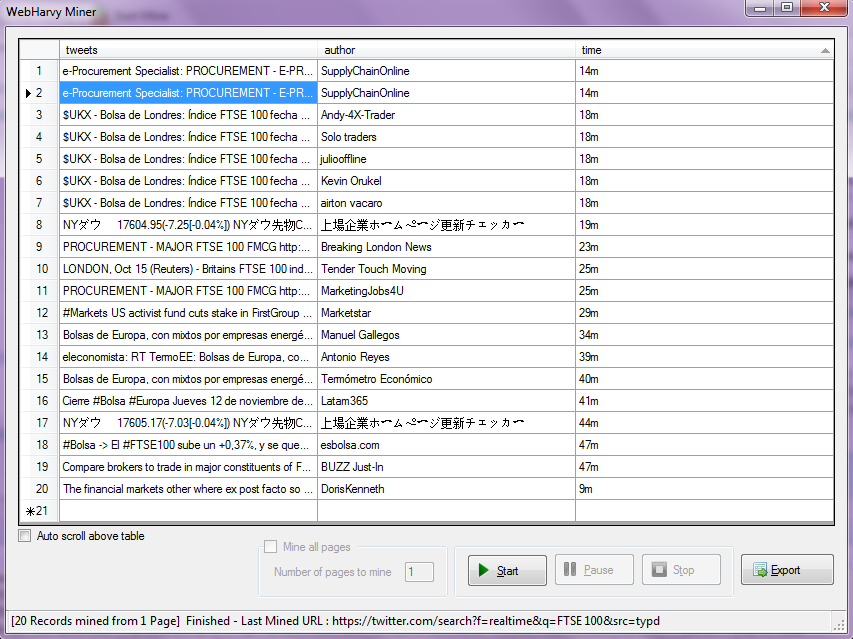


Figure 2. 7 Webharvey Miner Data

The miner data shown in Figure 2.7 can be saved in *.xlsx* format but this procedure is not free. It seems that the Webharvey can fit the requirements of Twitter extraction. However, there are drawbacks with Webharvey. The first problem is that the software cannot generate the exact post time of each tweet. As the Twitter post time is not completed in the search page, there is only the time from when the tweet was posted until now and this is shown in the third row in Figure 2.7; this will lead to chaos in Twitter sentiment procedure. Another drawback of Webharvey is that the software can only show the first page of twitter extraction results in miner data dialog, many tweets, include important ones, could easily be ignored. The third problem is that the Webharvey application has to be paid for and will increase the budget for this research.

1. R Programming Language

This section mainly focuses on retrieving text (tweets) from <https://twitter.com/> and analysing UK stock market tweets with the R word cloud function. In order to obtain tweets that include FTSE information, we need to do text mining on Twitter tweets containing the word “FTSE” using the Twitter API. Twitter authentication means creating an app at Twitter. Firstly, go to <https://apps.twitter.com/> and sign in with your Twitter account. Secondly, follow the instruction to name the application and give a brief description of your Twitter API. Twitter API also requires a valid URL for the website. The Twitter API Youchen\_SentimentAnalysis is shown in Figure 2.8 below.

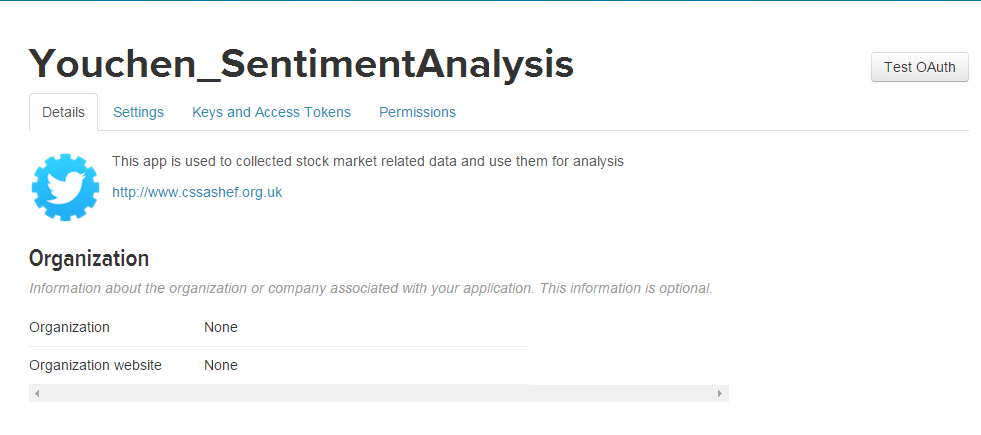


Figure 2. 8 Twitter API

Once the Twitter API is created, the developer will have a “consumer key”, “consumer secret”, “access token” and “access secret”. Researchers need to register these information to obtain the authority to extract tweets from Twitter.

### Web Mining and Twitter Sentiment Applications

1. Introduction

With the development of computer technology, the scale and amount of digitalized information has been greatly expended and enriched. Different platforms have made it possible to present the dataset of people’s daily life, such as the World Wide Web. Therefore, the development and popularization of the Internet will accelerate the development and dissemination of digitalized information.

The Web includes various kinds of information. As such, information research has a dilemma: information overload and information loss. The former means that large amounts of information can be difficult to analyse and process. Information loss means that it is hard to find specific data or information in the large dataset. Therefore, a methodology that is able to locate and analyse the specific data from Web information is necessary. Most Web information is stored as text or corpus form. Which means that the text data is the main storage form in Web information. Considering the scale and pattern of the information, it is essential to develop a web-based text data mining algorithm.

Web mining is a kind of data mining which includes web technology, data mining, text mining, natural language processing, artificial intelligence and other technologies that implement data mining algorithm into data science. Web mining is not only a tool for information retrieval, it will help to deal with the data extraction, analysis, modelling and predicting problem in the Internet. The flow chart of a web mining process is shown in Figure 2.9 below:

Web Data

Text Library

Result Validation

Text Feature

Figure 2. 9 Flow chart of Web data mining

1. Tweets’ Contents Mining from Twitter

According to Mao, Counts and Bollen (2011), measuring social sentiment is a challenging task in financial index prediction. Bollen, Mao and Zeng (2011) also state that “reliable, scalable and early assessments” of public online sentiment (Twitter sentiment/mood) in time scale is a key point to predict a financial market index. Twitter, as a social media network, is increasingly applied to share and exchange users’ opinions about different topics (Hu, et.al, 2013).

Traditionally, lexicon-based methods are a way to do sentiment analysis (Hu, et.al, 2013). A lexicon-based method determines the orientation for a document by calculating the overall sentiment polarity. Although lexicon-based methods have been widely used in text sentiment analyses, it is a challenge task for a lexicon method to determine the accurate sentiment polarity of tweets (Hu, et.al, 2013). Firstly, numerous tweets have insufficient information for researchers to evaluate the overall sentiment using a lexicon-based method. Tweets are different from reviews; each tweet has a limitation of 280 characters while a review is a thoughts’ collection. Secondly, most tweets include informal expressions, colloquial words and even abbreviations. Such popular expressions will make it difficult for Twitter sentiment analysis. Tweets are not as critical as reviews as a tweet always expresses one’s own thoughts in a simple way. However, tweets can still provide enough opinion-rich information for mining (Go, Bhayani and Huang, 2009). Thirdly, emoticons are widely used in Twitter. Many people use emoticons instead of words to express their feelings. Therefore, lexicon-based methods face numerous challenges when applied to Twitter mood analysis. Considering the limitations of lexicon-based methods, there is another algorithm that can provide more than 60% accuracy in Twitter sentiment analyses, even with emoticon data, and that is the machine learning method (Go, Bhayani and Huang, 2009). Although machine learning has advantages in short colloquial and informal expression tweets, it cannot deal with emoticons. Emoticons are considered as noise because they will influence the accuracy of machine learning algorithm (Go, Bhayani and Huang, 2009). In Go’s research, they strip out the emoticons to decrease the negative influence. Hence, the classifier will use the non-emoticon tweets to determine twitter sentiment (Go, Bhayani and Huang, 2009). However, some emoticons could be useful to test the Twitter mood and this is the limitation of the approach.

According to Bollen, Mao and Zeng (2011), tweets are able to express information about the sentiment level of their author. Often, there is much information related to sentiment in tweets. For example, one tweet “A **FTSE 100** Chief Executive now earns 120 times more than a full-time employee”, explicitly shows a negative viewpoint about the FTSE100. The sentiment of tweets is divided into positive, negative and neutral. According to Go, Bhayani and Huang (2009), if a tweet is a front-page news headline, Wikipedia words or other actual truth, it is considered as neutral. In Go, Bhayani and Huang’s (2009) research, they do not consider the neutral tweets, only positive and negative sentiment tweets. It was suggested that the Twitter mood information digitalized below could generate a better performance.

1. Twitter Data Pre-process

In recent years, online social network sites have increased rapidly around the world and the data/information technologies play a major role. The rapid development of the information and communication technology has already made it possible for data broadcast to be extremely critical (Hemalatha, Varma and Govardhan, 2012). Online social networks are a significant part of information exchange, sharing, communication and broadcast. One of the most popular social networking services, Twitter, is not only popular with young people, it is also widely used by older age groups. Twitter has changed many people’s lifestyle and it has wide range of applications such as “business development, reviews about various social activities and acceptance of any new ideas by means of Sentiment Analysis” (Hemalatha, Varma and Govardhan, 2012). These rich emotional data contain different information about public sentiment. To obtain these data, we need a social network service such as Twitter. Twitter is able to provide many real-time tweets from different perspectives. These reviews or comments towards a specific topic are generally given as positive or negative review or more accurately as sentiments, such anger, happiness or joy.

According to Uysal and Gunal (2014), common text pre-processing for sentiment analysis or text classification includes “tokenization, stop-word removal, lower case conversion and stemming”.

Tokenization represents the procedure of splitting a corpus into small units such as words and phrases or short sentences. “In other words, tokenization is a form of text segmentation and it is carried out considering only alphabetic or alphanumeric characters that are delimited by non-alphanumeric characters for example punctuations and whitespace” (Uysal and Gunal, 2014). According to Uysal and Gunal (2014), stop-words are those that are the most commonly used words in a language. Stop-words are usually irrelevant to the meaning of the text/corpus. Removing the stop words will help researchers to decrease the interference and will acquire the semantic information words.

Although uppercase and lowercase of words are the same in sentiment analysis, the lowercase transformation is widely used for text pre-processing. Converting uppercase to lower case will make the document-matrix clean and tidy. Fixed-prefix algorithm (Zemberek, 2013) and Stemming algorithm (Can et al., 2008) were applied in Uysal and Gunal’s research.

Petz, et. al,. (2012) believe that the following three text pre-process steps will help to acquire a satisfactory sentiment analysis result:

1. Splitting experiment text into short sentences or words
2. Acronyms, symbols and emoticons should be replaced
3. Stemming

For step one, texts are divided into short sentences or words, which is good for handling. For step two, researchers manually define the symbols and emoticons and replace these to make this information usable for sentiment analysis (Petz, et.al., 2012). The meaning of acronyms can be obtains from a dictionary that includes the most commonly used abbreviations (Petz, et.al., 2012). The stemming tool TreeTagger is used to deal with every single stem word (Schmid, 1995). Hence, the pre-processing of text includes re-structuring that is used to construct a text matrix, which can then be used as input for further analysis.

Considering this research, tweets from Twitter are the main source data of the research. Given the unique features of tweets’ data, the text pre-processing algorithm should meet the requirements of the Twitter data. According to Hemalatha, Varma and Govardhan (2012), data pre-processing should include the following main tasks:

* Removing URLS

“Generally, URLS do not make any contribute to the sentiment analysis in the informal text”. Removing the URLS will decrease interference and simplify the data.

* Filtering

Twitter users often use repeated letters to deepen the emotional expression of their feeling. These words will not be recognised as sentiment words by a computer and ignoring these will lead to loss of information. “The rule of the filtering is a letter could not be repeated more than three times”.

* Questions

There are some question words such as how, where, what, which that do not contribute to the sentiment classification and should be removed.

* Removing Special Characters

Tweets are an informal platform and Twitter users frequently prefer to use special characters in their Tweets. “If the special characters are not removed sometimes the special characters may concatenate with the words and make those words unavailable in the dictionary”.

* Removal of Retweets

Retweeting is the process of forwarding someone’s tweets to your Twitter friends. People who retweet another person’s tweets means he/she likes this tweet or agrees with the tweets.

Katariya and Chaudhari (2015) think that text pre-processing includes tokenization or normalization and this procedure can be divided into five operations which are “Lexical Analysis of Text, Stemming, Elimination of Stopwords, Index Terms Selection and Thesauri”. Katariya and Chaudhari also state that text mining is a technique to extract information from the documents. Text pre-processing plays an important role in sentiment analysis and the three most important pre-processing techniques are stop words’ removal, stemming and indexing.

### Twitter Sentiment Influence on Political Election

Twitter, as one of the most popular worldwide social network service will provide “an impressive amount of data about users and their interactions” (Harald, et al., 2013). At present, it is a popular research area to examine if and how online public services and online sentiment contribute to the world political issues. Honeycutt and Herring (2009) states that “previous research has suggested that microblog use that goes beyond the characterization of interesting novelty”. While, according to Jasen’s study (2009) the popularity and status of Twitter, appears to be an ideal candidate for online sentiment analysis.

Twitter has already provided a public platform for political communication and political debate. This platform makes use of the features of Twitter that include numerous comprehensive collection of big data that relate to the actual public issue. It seems that the US presidential election 2016 had provided a proper condition for Twitter content application. “Twitter is often understood as a derivative or miniature version of the regular blog and Twitter users share their updates to a network of followers. A user can follow any number of other users, although the user being followed does not necessarily have to follow back” (Larsson and Moe, 2012). And Twitter has been implemented as a significant predictor of the online political events (Gil et al., 2009), and it is the reasons for Twitter to become the most appropriate data source than any other online platforms. In order to study the Twitter sentiment, Java et al. (2007) has pointed out four general types of Twitter use:

* Twitter users who posts daily events and thoughts
* Conversations and communication by using the @ character
* Information sharing using URLs in their posts
* Twitter Users report latest news about current political and economic events

Meanwhile, Honey and Herring (2009) has implemented ground approach on the sample tweets and 12 distinct categories of tweets are founded, they are “about addressee, announce/advertise, exhort, information for others, information for self, metacommentary, media use, opinion, other’s experience, experience, solicit information and other”.

The United States of America's presidential election (or short for US presidential election 2016) has been a main political topic around the world. US presidential election 2016 will have a profound influence on the pattern of the world in economic political and military. There was a heated debate on who would win the presidential election, Donald Trump or Hillary Clinton? The discussion covers a wide variety of social media platform, such as News, Magazine, Forums and Social Networks (Twitter, Facebook).

Twitter is one of the most popular social media throughout the world, increasingly people will publish their opinion on every topic they are interested. Based on this, this research will mainly focus on the Twitter sentiment about these two political figures on tweets from 11/06/2017 to the election eve. In order to obtain the explicit Twitter sentiment, we would extract the Twitter about Donald Trump and Hillary Clinton separately and then the sentiment of each candidates will be analysed by Twitter lexicon based method and hybrid sentiment analysis model. The result illustrates the change of the public sentiment about the two presidential election candidates in the next chapter.

### Twitter Sentiment Influence on Stock Market Index

It is known that stock market price is an important indicator for the world economy. Based on behavioral finance theory, “stock market can be driven by emotions of market participants” (Nofer and Hinz, 2015). Because of mood information have been extracted from some Social Media such as Twitter to predict the stock market change (Nofer and Hinz, 2015), applying online sentiment data to predict stock market is becoming one of the most popular research area. Much evidence has shown that Twitter sentiment index is an important factor that can influence the world stock market price, or in other words, there is a relationship between Twitter sentiment and stock markets price (Bollen, Mao and Zeng, 2011) (Chen and Lazer, 2013) (Si et al., 2013) (Mao et al., 2012).

### Twitter Sentiment Influence on Brexit

Twitter, as one of the most popular social network platform has profoundly influence and changed people’s daily life. There are different kinds of topics is being talked about on every second. Tweets in Twitter always include a great deal of sentiment information that can easily affect peoples’ decision-making. This sentiment rich data is used to model and predict some social phenomenon such as voting. Whether the collective sentiment on UK Twitter users about Brexit are able to help to predict the result of upcoming referendum about United Kingdom withdrawal from the European Union (Brexit 2016).

The United Kingdom withdrawal from the European Union (Brexit 2016 or UK referendum 2016) has attracted the attention of the UK and the whole world. The Brexit will profoundly influence the pattern of the European Union and UK in political, economic and military. There is a heated debate on whether UK should withdrawal from the EU or not. In this part, researchers will mainly focus on discussing the feasibility of using Twitter sentiment to predict the UK referendum poll 2017.

Traditional methods that has been used to track and predict the polls is based on internet survey and telephone. Although the traditional methods can provide us an understanding of the voting situation, they are not able to cover all the information about some specific problem. There are certainly more data about Brexit on the various social networks such as Twitter.

Nowadays, modern people live in an information explosion world where people opinion, behaviour and activities will leave record on social networks such as Twitter. Social networks have profoundly influenced our daily life in every aspect. Numerous studies are about using the people’s comments in social network to predict specific information. Such as applying sentiment and search query to predict the box office of movie (Mao et.al, 2011), online sentiment predicts financial market (Mao et.al, 2011), internet search queries predict stock market volatility (Bordino et.al, 2012), Twitter data detect influenza epidemics (Aramaki et.al, 2011).

Why use Twitter? Twitter is one of the most popular social media network that allow users to post their opinion rich tweets on it. According to Bollen et.al (2011), the aggregate of large numbers of tweets at any post time may deliver a correct representation of people’s sentiment about some specific topic. Twitter has large amount of information about the topic about Brexit which can be acquired and this dataset is a key factor to analysis and predict. Compared with the Internet polls and Phone polls, Twitter will provide more data and more specific altitude about Brexit is included in their tweets for researchers to mining. However, Twitter data has its own deficiencies and problems. Firstly, tweets are not always tidy and sometimes include abbreviations and online expression. Secondly, users will sometimes post a figure to express their opinion and it is hard to analysis. Lastly, there many links, @ people which are irrelevant to our sentiment analysis. In next part, researchers will show how to pre-process the Twitter data.

### The Application of Twitter Sentiment Analysis

Twitter sentiment analysis has been widely applied in different research areas for monitoring and forecasting public sentiment (Jurek, Mulvenna and Bi, 2015). According to Mittal and Goel’s (2013) research, Twitter data is classified into different emotion index (happy, calm, kind and alert) and these indexes are user to predict the Dow Jones Industrial Average (DJIA) movements. Twitter data are also applied to find the correlation between movies box-office and Twitter feeds (Krauss et al., 2008). Grabner et al. (2012) applied a Twitter blog data in order to model the customer reviews of the hotels. According to Xu, Zhu and Bellmore’s research (2012), a novel text classification model was developed to recognise different emotions in Twitter posts. Hu, Wang and Kambhampati’s research (2013) states that Twitter sentiment is able to characterize event such as US presidential debate 2012 and President speech in 2011. Twitter sentiment can be also used to analysis the tourism threat (Garcia, Gaines and Linaza, 2012).

## Sentiment Analysis Methods

### Background and Introduction

Sentiment analysis and opinion mining, is a hot and popular research field that implementing public’s sentiment, emotion, opinion data towards substance like commodities, services, organizations, events, topic, products, individuals and their attributes. According to Liu (2012), sentiment analysis, also called “opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining” represents a large issue space. In this report, researchers will mainly use the term sentiment analysis. Sentiment analysis is a technique to distinguish positive, negative or neural sentiment towards specific subjects through textual information (Nasukawa and Yi, 2003). Normally, the algorithm of sentiment analysis requires much effective textual information. Liu’s research (2010) states that textual information can be categorized into two types: facts and opinions, objective expression of a products, events and their attributes are facts; “opinions are usually subjective expressions that describe people’s sentiments, appraisals or feelings toward entities, events and their properties.”

A few decades ago, when an individual wanted to buy commodities, he/she would normally seek opinions from his/her families and friends. When business organizations and companies wanted to find the public sentiment of some products and services, they mainly implement methods such as polls and surveys (Liu, 2012). With the explosive growth of World Wide Web users, especially with the development of Web 2.0 in the past few years that enables network terminal interaction. This will lead to there is large numbers of opinion textural information on the Internet. Considering about the research area, before the appearance of the World Wide Web, researchers mainly collect opinion data from survey, letters, the data insufficient and difficult to obtained. Therefore, the World Wide Web has fundamentally changed the information sharing and communication method. Opinion data in the Internet forums, Twitter, blogs, comments, discussion groups represents real time and massive data sources of textual information that can be implement in sentiment analysis. For an individual, if he/she want to buy something, instead of ask his/her friends and families he/she can read many online product reviews on the Internet. For business organisations and companies, they can obtain much opinion data about their products and service, and that is helpful for their decision making in the future.

It is still a difficult work to retrieve opinion textual data from the Internet because “there is a large number of diverse sources, and each source may also have a huge volume of opinionated text and text with opinions or sentiments” (Liu, 2010). Normally, opinions data are not directly shown in the corpus. Although it is not difficult for a human to understand the meaning of it, it is difficult for a machine to read, summarize and organize the online text information into a proper and usable forms. Therefore, an automated and real-time sentiment analysis algorithm is necessary. Sentiment analysis, also known as opinion mining, grows out of this need.

The sentiment analysis has been widely applied in social network (Twitter, Facebook) to acquire the public sentiment index (Go, Bhayani and Huang, 2009) (Kouloumpis, Wilson and Moore, 2011) (Bollen, Mao and Pepe, 2011) (Ortigosa, Martin and Carro, 2014) . Li and Wu (2010) have applied text mining and sentiment analysis for online data to detect and forecast hotspot. The hotspot semantic engine is able to automatically determine whether public sentiment about a company is positive or negative (Li and Wu, 2010); Fu, et.al., (2013) have designed topic model lexicon for the sentiment analysis on Chinese online social reviews; Greaves et.al., (2013) implement sentiment analysis method and online comments to capture the patient experience; Online data sentiment analysis can be also used to model product reviews (Jo and Oh, 2011) (Dang, Zhang and Chen, 2010). Above all, compared with other web platforms, Twitter usually have the more opinion information data and all kinds of topics. Twitter sentiment analysis have been widely applied.

### Twitter data Pre-process

As discussed in the previous part, when researchers want to retrieve data from social media platform (Twitter, Facebook and so on) using R program, there will be many problems to deal with. More specifically, R has some applied functions to retrieve data from Twitter.com. However, due to the development of Twitter, Twitter users tend to use a variety method to express their idea or feeling. Such as different languages, emoticons and abbreviation. Furthermore, every tweet may contain a variety of information, such as @ someone, links and graph. R seems to be able to display some of this information properly, however, sometimes it does not. Some information is unrelated to sentiment analysis or even have interference on experiment. Therefore, Twitter data pre-processing is very significant for sentiment analysis.

### Lexicon Based Method

Lexicon-based classification is defined as “a classification rule in which documents are assigned labels based on the count of words from lexicons associated with each label” (Taboada et al., 2011). Liu (2015) and Pang (2008) states that the Lexicon-based method is widely applied in academia and industry area, “with applications ranging from sentiment classification and opinion mining.” The resources for lexicon-based classifier, also referred to Sentiment Lexicon, is a collection of large numbers of words, word senses and phrases with their sentiment orientations (Ahire, 2014). The sentiment words are represented in several forms, such as positive or negative; more detailed strong positive, mildly positive, strong negative; index value from -1 to +1 (Ahire, 2014). According to Mohammad (2013), “the NRC Lexicon is a list of English words and their associations with eight base emotions (anger, fear, anticipation, trust, surprise, sadness, joy and disgust) and two sentiments (positive and negative)”. Although there are rich varieties of sentiment lexicons, lexicon based method still have some defects: 1. the application condition of different sentiment lexicons is unclear; 2. Sentiment dictionary cannot contain all the English words which makes the lexicon incomplete; 3. Some multi-words meaning will be ignored (Eisenstein, 2017).

### Text Mining

Text mining, also referred to text data mining, can be regarded “as going beyond information access to further help users analyse and digest information and facilitate decision making” (Aggarwal and Zhai, 2012). Text mining and data mining are not distinct concepts, they are all based on the past examples (Weiss et al., 2010). Although the learning methods of text/data mining are similar, the composition of example is different (Weiss et al., 2010). Because of majority of (80%) information online is stored in text form, research about text mining is considered to have huge commercial value (Gupta and Lehal, 2009). Gupta and Lehal (2009) also states that “text mining is a young interdisciplinary field which draws on information retrieval, data mining, machine learning, statistics and computational linguistics”. In 2.4.5, we will introduce and compare some machine learning methods for text mining.

### Machine Learning Methods for Document Classification

“Document classification is a growing interest in the research of text mining” (Ting, IP and Tsang, 2011). The objective document classification is to assign a text/corpus into appropriate categories or classes (Weiss et al., 2010). With the rapid increasing of the text data, especially the explosive growth of internet text information. The document classification has been applied in spam email filtering, email classification and website categorization (Ting, IP and Tang, 2011). Because of it is impossible to label all the document categories manually. Therefore, the data mining methods such as support vector machine (SVM), Naïve Bayes (NB), k-nearest-neighbour (KNN), Artificial Neural Network (ANN) and decision tree are developed to implement in document classification problems (Ting, IP and Tang, 2011). In this part, researchers will review and compare these algorithms.

1. Support Vector Machine (SVM)

Support vector machine (SVM) is one of the supervised learning technique that include many advantages in document/text classification (Moraes, Valiati and Neto, 2013). Based on studies, SVM is one of the “discriminative classification methods which are commonly recognized to be more accurate” (Khan et al., 2010). Numerous research show that SVM classifier has been applied to social network data classification: Go, Bhayani and Huang (2009), Jiang et al., (2011) and Wang et al., (2011) highlight the significance of SVM for Twitter sentiment classification; SVM has also been used to classify the Twitter news (Dilrukshi and Zoysa, 2013); Singh et al., (2013) implement SVM methods for movie reviews’ sentiment analysis. Although SVM has been widely applied into document/classification, the property of SVM show less affected by noise and the training time is usually short (Moraes, Valiati and Neto, 2013). Moraes, Valiati and Neto (2013) have indicated some drawbacks of SVM: 1. Other machine learning method ANN significantly outperform SVM in classification accuracy of unbalanced data; 2. ANN also performed better than SVM in the context of balanced data.

1. Context-Sensitive Learning Methods

Context-Sensitive learning methods are first proposer by Cohen and Singer (1999). By using context-sensitive learning methods, the training set will be simplified and the efficiency of the training and classification process have been improved. However, the drawback of this method is the classification accuracy is depended on the feature distribution. When the feature is well distributed or the feature boundary is clear, the classification accuracy is acceptable; when the feature is fuzzy, the classification result is not acceptable.

1. KNN Algorithm

KNN classifier is based on the vector space model, the content of test document is formalised into the vector in the space model. The similarity of a test document data and the training dataset will be measured, the category will be determined by calculating the weighted distance. KNN classification algorithm is with high accuracy and robustness, it also performed well in non-normal distribution data. Although KNN is a superior algorithm in classification problem, KNN still have its deficiencies:

* Processing high dimension text dataset will lead to high complexity of model itself.
* When a new sample is going to be processed, the distance between the new sample and the training data (similarity) have to be measured. Hence the effectiveness of KNN will be reduced.

Considering about the deficiencies of traditional KNN algorithm, there exit many improved KNN algorithms. There are two main ways to reduce the computational complexity and improve the efficiency of the algorithm: 1. Applying the dimension reduction for the high dimension text data. For instance, Dumais (2004) applied latent semantic analysis (LSA) to reduce the dimension of text data; feature vector aggregation can also help to reduce text data dimension (Li, et al., 2012); Based on Qiu et al., (2012) research, feature extraction method has been applied to KNN. 2. There is another method that is able to achieve the text dimension reduction by using small sample to classification rather than original dataset. More specifically, the sample is chosen from the original text dataset to become the new training sample, or delete some data in the original dataset and the rest sample will become the new training sample data. The algorithm Condensing, Editing, MultiEdit and BC-iDistance can help us obtain the new training sample.

In conclusion, KNN is an easy and effective algorithm in text classification and it is widely used in text classification problem. Currently, the research of KNN algorithm mainly focus on the feature dimension reduction and sample data cut. With the development of KNN in recent years, there exist many superior KNN improvement algorithms. Based on the Projection Pursuit (PP) and iDistance, a new text classification method PKNN is proposed. The core of this algorithm is the n1 selection. Large n1 will lead to low classification efficiency; small n1 will lead to low classification accuracy.

1. Naïve Bayes Methods

Naïve Bayes classifier is a probabilistic classifier based on Bayes’ theorem. Although Bayes classifier algorithm is simple, the model classification output is usually effective and accurate especially in big data base like Decision Tree and ANN algorithms. Narayanan, Arora and Bhatia (2013) states that Naïve Bayes classifier has the following characteristic: the algorithm is simple to implement; with a high classification accuracy; high working efficiency. Therefore, NB classifier performs well on the classification of text data and digital data. However, the features in NB classifier have to be independent with each other, hence the test dataset should satisfy the independence assumption to obtain a precise classification result.

1. Artificial Neural Network

According to the features of the Artificial Neural Network (ANN) that were discussed before, it is known that multi-layer ANN model can be used in complex, nonlinear dynamic system. Recently, the ANN have been implemented in text mining, especially in text classification problems. Sometimes, the model outputs are very different from the actual results, the ANN model will adjust the model itself to meet the requirements which is called self-adaptive. ANN are robust and any operation will not influence the overall model outputs. Paralleling processing is the main characteristic of ANN, either calculation speed or data storage can be guaranteed.

1. Decision Tree

Decision tree algorithm is a tree-shaped model or graph of decisions and all the consequence and their probabilities. This algorithm is able to generate proper classification rules from a large complex, random dataset. The famous decision tree method ID3 is first introduced by Quianlan in 1986, and then, in order to meet the requirement of big dataset, SLIQ and SPRINT is proposed Apte, Damerau and Weiss, 1998). Generally, there are many features of a corpus, applying this method will largely increase the complexity of decision tree model.

1. Model Classification Performance Validation

The basic assessment principle of text classification is output accuracy and model complexity, which means researchers need to find an equilibrium between classification accuracy and complexity. Therefore, there are two important factors in text classification model evaluation: Precision and Recall and the relation is shown in table 2.1 below:

Table 2. 1 Relationship of classification evaluation

|  |  |  |
| --- | --- | --- |
|  | Condition Positive | Condition Negative |
| Algorithm determined Positive | A | B |
| Algorithm determined Negative | C | D |

According to Table 2.1 above, A represents the condition is positive and algorithm result is positive (true positive); B represents condition is negative while the algorithm result is positive (false positive); C represents the condition is positive while the algorithm result is negative (false negative); D represents the condition is negative and the algorithm result is also negative (true negative). The Precision and Recall are:

(2.3)

(2.4)

The precision and recall should not be considered independently, and there is another assessment method F1 test which is shown below:

(2.5)

1. Conclusion

The 2.4.5 reviews different machine learning methods that applied in document sentiment classification. Different machine learning methods (Bayes, SVM, KNN, ANN, Decision Trees) have been introduced and compared with each other in detail. With the development in document classification, some hybrid methods/algorithms have been proposed and evaluated (Yang et al., 2016) (Tang, Qin and Liu, 2015) (Gayathri and Marimuthu, 2013). However, NB, SVM and KNN classifier have shown better classification results among these algorithms (Khan, 2010).

### How the Machine Learning Algorithm Affects this Research?

The machine learning algorithm is an effective and powerful method for predictive models fitting and large scale, un-stationary and high-dimensional dataset classification. Research studies have shown that the machine learning method are not only used in engineering and computer science, but also increasingly used in economics. Considering about the economic research and characteristics of economic data, more machine learning methods will be more used in the future research.

As discussed above, the machine learning algorithm is used to find the sentiment level of Twitter feeds. In big data world, Supervised and Unsupervised machine learning is fundamentally about classification and prediction, different kinds of information such as Twitter sentiment, Financial index can have a potential impact on the outcome. In this process, machine learning plays a role in filter and classification.

Machine learning is used to deal with the problem: How these factors x can be used to predict another factor y? Which of these individuals belong to which class? The results shown that the machine learning algorithm is an effective and powerful ways in these problems. In this big dataset, some the data is used to build a training model and the remaining data is used to measure the predictive power of the training model. It is important to choose the appropriate training dataset and testing dataset, or it will affect the accuracy of prediction and classification result. Aki Information Criteria and Bayes Information Criteria are widely used to classify the training and testing part.

The machine learning techniques are immediately applied to stock market Tweets sentiment analysis that the sentiment index is important to predict stock market variety. Researchers are beginning to apply these techniques on classifying the sentiment polarity of tweets. In order to have an appropriate prediction of Tweets, large numbers of training dataset are needed and this will need much manual work.

This research focuses on modifying the machine learning methods in order to find effective and efficient ways to obtain the Twitter sentiment. Once the reliable Twitter sentiment is acquired, the Twitter Sentiment index is really significant in either modelling the stock market system or predicting the Stock market index. Using the combination of Twitter sentiment and machine learning methods researchers can estimate the daily stock market index change.

## Social Networks and Complex Network

### Introduction

In the big data study area, big data has a profound impact on people’s daily life. In order to study how big data affects people, Social Network is a method to help us to understand and visualize the online big data.

How computer virus spread on the Internet? How stock market index, such as DJIA, FTSE and world political issues influence the world economy? How news affects public sentiment and public cognition. How the infectious diseases, such as flu spread among human beings and animals? How individual behaviours on social media platform such as Twitter, Facebook and YouTube affect the public. Although these problems above are different from each other, each problem is related to Social Networks research. Recently, research has shown that social networks can be widely applied in different areas. According to Wang, Li and Chen (2006) Complex network, Internet, social network, Economic networks, Transportation network and Neutral network have many similarities. This part of the literature review will discuss the concept and application of social networks.

### Complex Network

Before discussing Social Networks, we should know their basic concept. Different kinds of social networks exist in the real world. Typical network was consisted by nodes and connections, nodes represented different individuals and connections represented the relation between different individuals (Zhou et al, 2004). Estrada (2011) states that “A network (graph) is a diagrammatic representation of a systems, it consists of node (vertices), which represent the entities of the system. Pairs of nodes are joined by links (edges), which represent a particular kind of interconnection between those entities.” According to Zhou (2004), the Social Network is a network that have large, real, complex system topological features. Furthermore, complex network is more complex than mesh graph and random graph.

### Complex Network Properties

This part will discuss some basic properties conception of Complex Network. Before discussing the Complex Network, some basic terminology needs to be deliberated:

* Nodes: Nodes are usually some points that represent individual, group or object. For instance, in Twitter Network all nodes represent Tweet message posted by Twitter user.
* Links: Links represents some relationship between two nodes. For example, links in Twitter network represent the relationship between Tweet message and its author.
* Network: Network is a gathering of different nodes and the nodes’ relation with each other, networks are usually applied to represent real world system.
* Degree of node: The degree of a node means the number of links that connected to the node.
* The average path length: In the networks, the number of shortest distance of two nodes; the average path length is the average distance of all the pairs’ length.
* Clustering coefficient: Clustering coefficient is the degree of nodes in the network tend to cluster together.

There are some differences between regular network and random network. According to Zhou (2005), the value of the average path length and the clustering coefficient of regular network are large, while the value of the average path length and the clustering coefficient of random network are small. Complex Networks have many unique statistical properties, the most important properties are small-world effects and scale-free properties (Zhou, 2005, p32).

### Social Network

Social Network analysis (SNA) is the process of investigating social structure through the use of network and graph theories. More specifically, SNA applying the nodes to represent individual actors, people, locations or other things in the network and the ties to represent their relationships or interaction that connect them.

1. Public Media and Social Media

Social media is a kind of social network that users can interact with others through this platform, such as information exchange or share something they find is interesting. There are different types of social media, for instance, newspapers, blogs and micro-blogs, Twitter, Facebook and YouTube. Because of the information technology growing rapidly and the development of Internet, more Internet social media such as Twitter, Facebook and YouTube has already implemented it for users to interact with other users at the same time (Mangold and Faulds, 2009). Evidence have shown that social media are becoming more popular. According to Lewis, Purcell, Smith and Zichuhr (2010), 73% of American teenagers use social media website, such as Twitter and Facebook on September 2009, the statistic number continued to rise from 2008. With the rapid development of Internet, “online social media describes a variety of new sources of online information that are created, initiated, circulated and used by consumers intent on educating each other about product, brands, services, personalities and issues.” (Blackshaw and Nazzaro, 2004, p2). Nowadays, Social media includes websites, blogs, social networks, email groups and so on. According to Mangold and Faulds (2009), compared with different social media, Twitter is one of the most popular social media network and researchers could acquire much useful information from it. Twitter information and social network theory has been used to “understanding global spread of disease” (Brennan, Sadilek and Kautz, 2013) and “modelling spread of disease from social interactions” (Sadilek, Kautz and Silenzio, 2012). Furthermore, this theory has also been applied to understand the social behaviour, such as mining and analysing the Twitter data during Australian flood 2010 (Cheong. F and Cheong. C, 2011), analysing the eating behaviour of US youth (Corrado and Distante, 2012). The following part will show some research of social networks, Twitter.

1. The Spread of Epidemic Disease

Traditionally, in order to monitor the spread of epidemic disease, the only method is to obtain the disease data that from the doctor record of hospital and health service (Sadilek, Kautz and Silenzio, 2012). This method is inefficient to acquire the epidemic trend and would leads to the overspread of disease. More specifically, the doctor record could only be acquired after and affected person go to see a doctor. Instead of go to hospital, some affected people sometimes choose to search relative website of consult their family or friends. Therefore, “Monitoring and forecast of global spread of infectious diseases is difficult, mainly due to lack of fine-grained and timely data” (Brennan, Sadilek and Kautza, 2013). In this situation, researchers can rarely obtain the exact information.

With the development of digital media and social network, social media network “has been successfully used to significantly reduce the latency and improve the overall effectiveness of public health monitoring” (Sadilek, Kautz and Silenzio, 2012). For instance, Google Flu Trends can model the epidemic flu using “geo-located search queries” (Ginsberg et al, 2008).

Speaking of Twitter, it is a popular social media networking service that enables registered users to write no more than 280 characters messages. Twitter support mobile devices to become terminals. Hence, tweets posted from mobile phones and always have accurate location (Brennan, 2013). Users can follow other users on Twitter. When two users follow each other, the two users become friends to each other. Based on Sadilek’s (2012) research, the experiment data are extracted from Twitter message and they identified the Tweets that show the author is ill. All these classifier processes are finished by a support vector machine (SVM) classifier. After identifying the affected individual, researchers collected the geo-tagged tweets which can easily predict the susceptible populations (Brennan, et al, 2013).

Although the information is time varying and we know people who become sick once he/she posted it on Twitter, our observation still cannot cover the sick people who do not post their feelings on Twitter. Although Twitter is popular, there are still people who do not use Twitter to express their feelings. Both of these mean the number of infected people who we have observed is smaller than in the real world.

### Complex/Social Network Platform

Complex Network and Social Networks include many nodes and links that has been found in many different area, for instance, “fenealogies, flow graphs of programs, molecule, computer networks, transportation networks, social networks, intra/inter organizational networks” (Batagelj and Mrvar, 2009). Recently, the number of complex network analysis software is growing rapidly. However, according to Batagelj and Mrvar (2009), some algorithms of Complex/Social Networks are unsuitable for analysis networks. For the widely used software, the difference of these is the data analysis ability and the compute speed (Hu and Zhu, 2010, p33). For instance, *Igraph* can deal with millions of data point, however, the limitation of *UCINET* is 30000 data points (Hu and Zhu, 2010). Another different is the function if information visualization. *Pajek* and *NetMiner* could deal with the information visualization process. There is about 70% of the Complex/Social network analysis software have such function (Wang, 2009, p96). Information visualization could show the structure of complex network and it helped us mining the information inside the network. Then, we would compare different Complex/Social network software Pajek and *UCINET*.

* Pajek

Pajek is a kind of software that used for windows to analysis large networks (Batagelj and Mrvar, 2009). Pajek could process over 1 million nodes network (Hu and Zhu, 2010). Pajek is a program with information visualization function. Furthermore, Pajek can do the clustering analysis and show the relationship of different cluster (Hu and Zhu, 2010). So Pajek has been used for large scale networks.

* UCINET

UCINET is a kind of software that used to analysis small networks. UCINET data are all stored in matrix format and can process at most of 32767 nodes network (Hu and Zhu, 2010). UCINET can read excel and other software data to make it convenient to complete data exchange process (Wang, 2009).

* NodeXL

NodeXL is a network visualization and analysis software based on Microsoft Excel 2007-2016. The NodeXL can also access social media network data and network matrics. The strong point of NodeXL is: 1. It is designed for the users who have limited programming knowledge. 2. NodeXL is able to import data (network figure) from UCINET and Pajek.

In conclusion, there is a wide range of complex network software. Considering about the practicality, Pajek, UCINET and NodeXL can be applied in the future research. For small scale networks, UCINET works well than Pajek. However, Pajek is more suitable for dealing with large scale networks. Compared with UCINET and Pajek, NodeXL is designed for the users who with poor or even no programming skill to help them to extract, analysis and visualize social network data (Bonsignore, 2009). As discussed above, NodeXL is embedded in to Microsoft 2007 to 2013, a variety of graph format data, matrices can be easily imported into this worksheet. Furthermore, NodeXl allows users to collect network data from online social media platforms. Lastly, NodeXl provide an efficient platform for graph analysis and graph visualization. Hence, in this research, NodeXl is the best platform for social network analysis.

## Conclusion

With the development and the applications of big data technologies, implementing social network data to model real world problems such political election, referendum, stock market change and crude oil price has already becoming a hot research area. In this process, data mining, text mining, text/document classification, sentiment analysis/opinion mining, system identification techniques about complex nonlinear systems and wavelet analysis will be studied and implemented. We have found that current data mining/extraction software or methods for Twitter are usually chargeable, what’s more, the extracted text datasets are not update and real-time. In order to model the political and economic variety, the Twitter text information is required to be real-time and updated. Compared with other sentiment lexicon, NRC lexicon is able to distinguish text sentiment into positive, negative and neural and text emotion into anger, anticipation, fear, joy, disgust, trust, surprise and fear. The applications of NRC to investigate the public opinion on Twitter about political and economic issue is innovation. Based on NRC lexicon, we will find which detailed sentiment or emotion contributes to the real-world problems. Machine learning techniques about text/document classification have been applied and studied many times. However, considering about the features in Twitter, there still have no appropriate models or algorithms for Twitter contents classification. Lastly, although system identification nonlinear models have been proven effective in different area, the most widely used algorithms in dealing with severely nonlinear and un-stationary systems are artificial neural network (ANN) and ANN related methods. Whether the wavelet based nonlinear models are able to reflect the nonlinearity of the complex system such as stock market price will be investigated.

The results of the literature reviews have shown that some severely nonlinear and un-stationary system such as stock market can be modelled and predicted with some specific nonlinear models. Such as Artificial Neural Network (ANN), Nonlinear regression models, 2nd order Volterra models and Wavelets. As we discussed in 2.3, there is a connection between Stock Market index variety and economic variables, social events and public sentiment. In this chapter, my literature review finding demonstrates that although there are some mathematical methods of nonlinear models are able to offer an acceptance predictive power for the stock market system, it is still difficult to obtain a reliable and profitable model for the stock market process. It is known that stock market changes are greatly influenced by global economic, investment sentiment and political events. For current research, linear models, nonlinear models, neural networks have been applied for stock market price/variety modelling in order to find the features of the stock market change. The complexity of stock market system will make it difficult to obtain a perfect model for stock market predictions. As discussed above, wavelet NARX model with orthogonal least square (OLS) algorithm has its advantages in dealing with severely nonlinear and un-stationary process, and the detailed discussion and description will be presented in chapter 5. The shortages and limitations of these existing models have made it possible for wavelet based NARX model with OLS to be applied to model and predict the nonlinear and un-stationary system such as UK Stock market price, FTSE100.

According to Bollen, Mao and Zeng’s (2011) research, Twitter sentiment has the power to influence the Stock Market change/variety, however, how these indexes influence the stock market price is still unknown. In order to explore the correlation between Twitter sentiment and stock market variety, the first step is developing a program for extracting text information online. As we discussed in 2.3.5, compared with current online data extraction method, Twitter API based in R language for mining (extract and analysis) real-time and information-rich Twitter data from the Internet will be used. There is one disadvantage of Twitter API based in R is the extracted tweets include different kinds of worthless/disturbance information and some of them may interference the sentiment analysis results in step 2. Therefore, how to tidy the messy Twitter information extracted by Twitter API R to reduce the interference from the irrelevant Twitter information and mining the effective data from Twitter have become the significant problems that need to be solved. Studies about text data pre-process include: removing URLS, filtering, remove special characters, removing retweets. We will implement R language and its Nature Language Process (NLP) functions to pre-process the extracted Tweets in chapter 3 and chapter 4.

Additionally, we have reviewed some sentiment analysis methods and document/text classification methods, such as lexicon based methods (NRC, AFINN), machine learning methods (NB, KNN, SVM), the advantages and disadvantages of each method is compared and evaluated in 2.4.5. Considering about the classification methods about text data and our experiments needs, NRC lexicon method will be applied. According to Khan (2010), KNN and NB outperform than other machine learning methods in short text/document classification problems. Therefore, these two algorithms are chosen to conduct our experiments. The literature review results also show that several hybrid or combination of algorithms with feature selection techniques are shown appropriate performance (Khan et al., 2010). By using these text/document sentiment classification methods, Twitter FTSE data, Brexit 2016 data and US presidential 2016 Twitter data will be investigated in this research. However, the Twitter sentiment model still have its limitation and drawbacks. Twitter has experienced a rapid development in recent years, with a large increase in the number of users, the content discussed on Twitter covers all the popular topics of all walks. Despite this, there are still large number of people who prefer not to express their opinions on Twitter, this will lead to the limitation of Twitter sentiment results. Or some users think Twitter is an emotional platform, so they will put perceptual and impulse Tweets rather than rational and thoughtful tweets. This will cause deviation of the analysed results. Lastly, there are some political and economic groups will put some Tweets based on their own interest and this behaviour will affect out models’ judgement to the real Twitter public sentiment.

Based on the behavioural economics theories, public sentiment is a significant factor that is able to influence the investors’ decision on making their decisions and investments. Research has also shown that Twitter is a platform that includes the information about public sentiment data and this real-time dataset can be applied to predict stock market. Applying the Twitter sentiment indexes that acquired from the lexicon and machine learning sentiment models. Different sentiment and emotion indexes will reveal different predictive power in UK stock market power. Therefore, the predictive power of our stock market system is believed to improve by applying Twitter sentiment indexes to the Wavelet OLS models.

# Chapter 3. Sentiment Analysis for Web Information

## Introduction

With the development of the Internet technology, especially the popularity of Web 2.0, there are large numbers of Internet users have already achieved a change from Internet information acquirer to Internet information maker. Twitter, as one of the product in Web 2.0 period, has already experienced a user’s explosive growth. Twitter users can post their tweets from their own terminals (PC or mobile phone) anytime, anywhere. Up to now, there are more than 319 000 000 active users on Twitter and 900 000 000 tweets are posted a day. In these large amounts of text information, emotional text information accounts for a large proportion. Twitter sentiment analysis implements machine learning algorithms and lexicon method to mining and organise tweets, in order to recognise the sentiment and mood of the Twitter.

Currently, the Twitter has already becoming one of the world most popular APP. Users are implementing Twitter to express their opinions and views anytime, anywhere. These topics related to politic events (Brexit 2016, Terrorism and US presidential election), social hotspots, economic issues (world stock market price, exchange rate and oil price), technology, travelling and shopping. Because of different people will make different opinions and views based on his/her cognitions and the cognitions is based on the family, education and professions. Hence, the view and opinion are showing a large difference. By means of Twitter, these differences are very likely to spread explosively, and it will have an influence on social, politic and economic.

Sentiment analysis has already been implemented in different kinds of non-security research domains for modelling and forecasting public sentiments. There is a sentiment analysis algorithm that based on the semantics research which is represented by lexicon based method

* Lexicon based method

Lexicon based method about semantics of sentiment analysis are proved to be robust, the experiment result performed good in cross-domain, and can be easily enhanced with other knowledge (Taboaba, Brooke and Stede, 2009). Furthermore, lexicon based method on sentiment analysis has been proved well performance on online blog postings “without any need for further development or training” (Murray et al., 2008). Twitter sentiment

In this part, researchers will implement Lexicon Based method on Twitter sentiment analysis. The structure of this part is: In 3.3, research focus on finding a fast and effective way to extract some specific tweets on Twitter. Next section, research will implement some method on Twitter pre-process in order to improve the accuracy of the sentiment analysis. Next, some case studies about Brexit 2016, US presidential election and UK stock market price will be applied for the sentiment analysis and different algorithms will be used for the sentiment classification, result will be evaluated and compared.

## The Significance of Twitter Information

Twitter has been constantly changing our way of life. Increasingly people choose Twitter as the platform to post their comments about political, economic and entertainment topics. Compared with traditional forum service, although Twitter can only implement the information interaction within their own friend circle on Twitter, Twitter still win the customer based on its convenience. With the big data age comes, there are large sentiment-rich datasets are contained in tweets and mining these tweets can provide much useful information for sentiment analysis and opinion mining. Thus, whether this sentiment rich data can be used to model and even predict some political and economic problems, such as voting and stock market prices.

## How to Extract Tweets on Twitter

### Twitter extraction with R

In order to extract text from Twitter with R, we need R packages “twitteR”, “RCurl” and “tm” to let it run successfully. By applying “consumer key”, “consumer secret”, “access token” and “access secret” the created by Twitter API, we can use searchTwitter function to retrieving Tweets containing “FTSE” from Twitter. The result is shown in Figure 3.1 below.

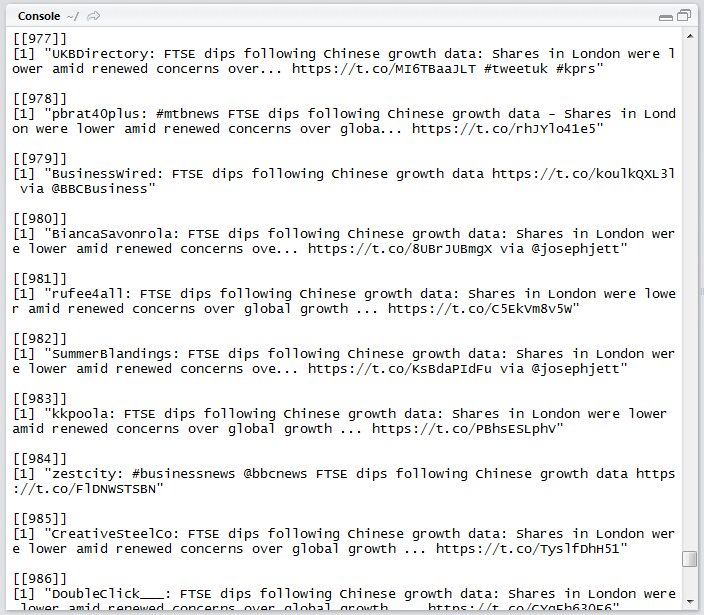


Figure 3. 1 Retrieving Tweets Results

As shown in Figure 3.1 above, 1000 tweets containing the word “FTSE” were obtained. By using this method with R, we can extract useful tweets from Twitter to help our analysing and forecasting work.

### FTSE Twitter Word Cloud

A Word Cloud is a visualization of document or text in terms of the words frequency. In other words, the more frequent a word, the larger its size is. Speaking of FTSE tweets, the word cloud will help us determine the most influential words in a day and it will help us to determine the public sentiment of that day. By applying “twitteR”, “RCurl”, “tm” and “wordcloud” packages, the word cloud of FTSE’s Twitter is shown in Figure 3.2 below.



Figure 3. 2 FTSE Word Cloud

It can be seen from Figure 3.2, because of some emotion words in the word cloud, the author has obtained a brief overview of public sentiment from today’s FTSE. According to Figure 3.2, the Concerns is the biggest words in the word cloud, which means that the words had been most frequently used in recent tweets about FTSE. Therefore, author will think that the public will hold a negative sentiment about recent UK stock market.

In the future research, the geographical location of the Twitter users will be considered. By applying Twitter users’ geographical location, we can obtain more specific public sentiment data relating to the UK stock market.

## Twitter Data Pre-process

When researchers want to retrieved data from social media (Twitter, Facebook and so on) using R program, there will be many problems to deal with. More specifically, R has some applied functions to retrieve data from *Twitter.com*. However, due to the development of Twitter, Twitter users tend to use a variety method to express their idea or feeling. Such as different languages, emoticons and abbreviation. Additionally, Tweets may contain a variety of information, such as @ someone, links and graphs. R seems to be able to display some of this information properly, however, sometimes it does not. Some information is unrelated to sentiment analysis or even have interference on experiment. Therefore, Twitter data pre-processing is highly significant for sentiment analysis.

This part will undertake a case study of Donald Trump’s tweets about a recent released movie “Captain American: Civil War”. Researchers use “Captain American” as the key words to retrieve recent relevant tweets data by R, some of the results are shown in Figure 3.3 below.

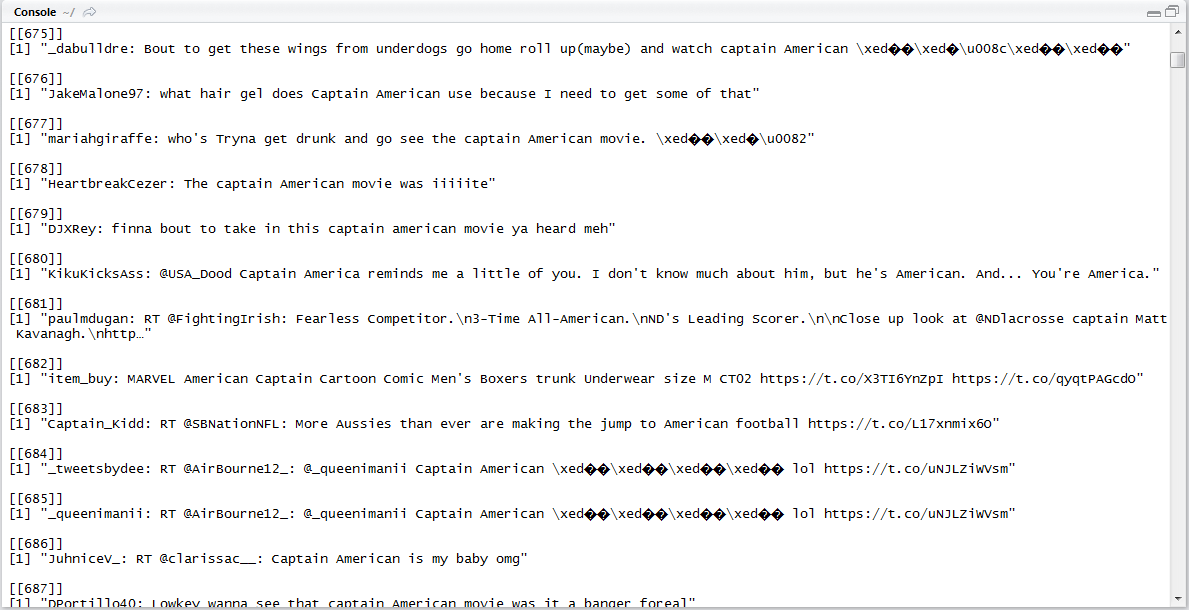


Figure 3. 3 Unprocessed Tweets

It can be seen from Figure 3.3 above, some of the information in tweets is not required. For example, the author name, @, http links and. Although R can show the emoticon property, it is still need special method to process that will be discussed later.

By removing @ people, http links and graphs, the processed tweets are shown in Figure 3.4 below

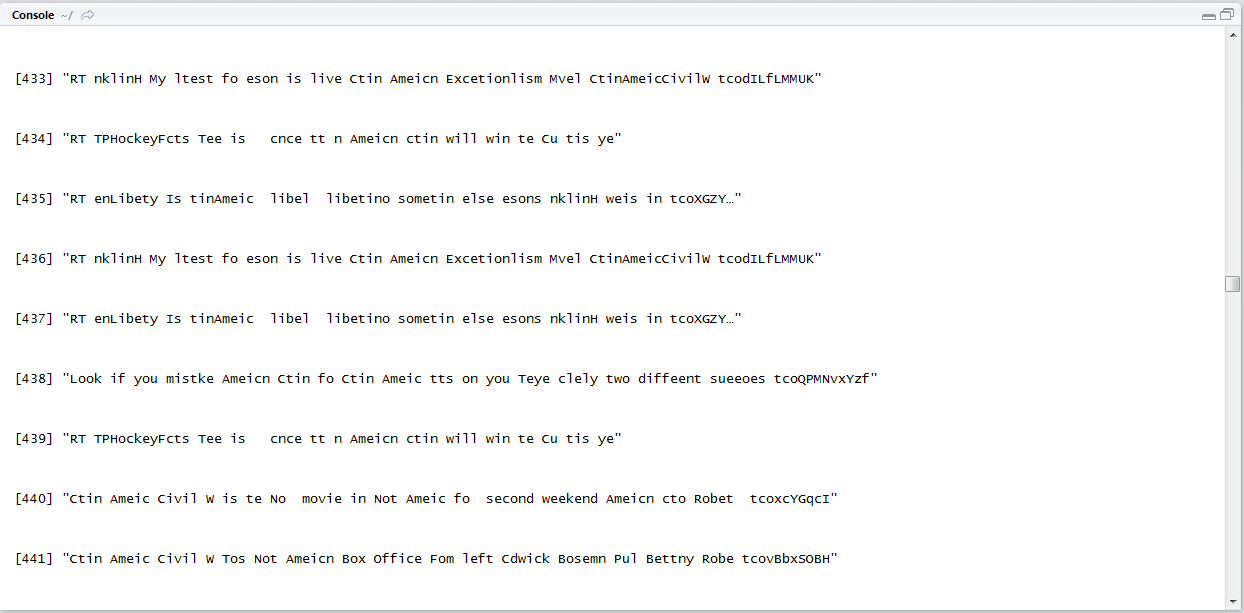


Figure 3. 4 Pre-Processed Tweets

As shown in Figure 3.4 above, without @, links and author, the tweets look more clean and tidy than before. Although the processed data is better and can be used to sentiment analysis, there still have 3 main defects: 1. There exit some unknown letters. 2. Some garbled letters appear. 3. Many words missing one or two letters. 4. There are some different languages exit on the Twitter and the Twitter can only recognise English.

Considering about the 4 defects of the pre-processed Twitter contents, there are still numerous interfaces and unrecognised languages. As Twitter is a public worldwide online social platform that every people in the world can use it. Some of the Twitter users prefer use their own language or combine their own language and English rather than only using English to express their thought. In order to deal with this situation, researchers will extract all the English content in Twitter contents. Here is a case study of US presidential election 2016, researchers extract one day’s tweets about Donald Trump which is shown in Figure 3.5 below.

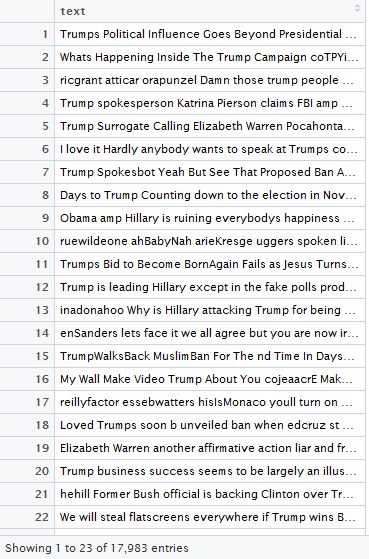


Figure 3. 5 Sample Donald Trump’s Tweets

As can be seen from Figure 3.5 above, there are a total of 17983 Tweets about Donald Trump. This text content contains different kinds of language data, in order to decrease the interface of different languages, researchers applied some lexicons to extract the English words of this text content and the result is shown in Figure 3.6.

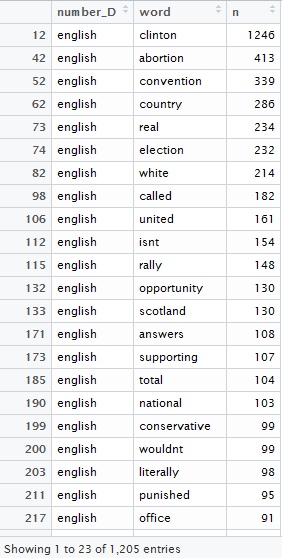


Figure 3. 6 Sample Twitter Word Frequency of Donald Trump

Figure 3.6 shows one day’s Twitter word frequency of Donald Trump. The first column is the word number, the second column is the language type, the third column is the English words in the text content and the last column is the number of the occurrence of the English words. Based on this, we could generate a word cloud which is shown in Figure 3.7 below:

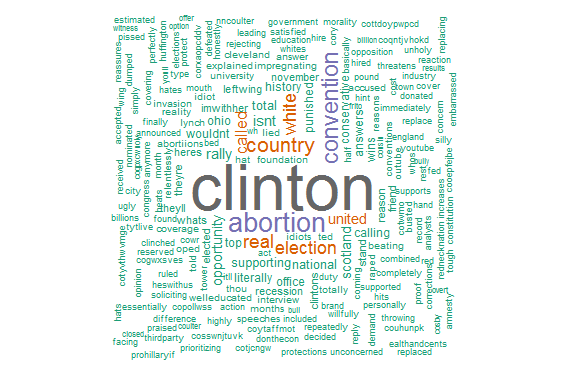


Figure 3. 7 Sample Twitter Word Cloud of Donald Trump

## Sentiment Analysis for Twitter

### Introduction

The goal and the task of the lexicon based sentiment analysis of Twitter text data is: 1. Identify the sentiment polarity of Twitter (Positive, Negative or Neural). 2. Determine the proportion of different emotions in the Twitter contents (anger, fear, anticipation, trust, surprise, sadness, joy and disgust). Noteworthy, the Neural sentiment cannot be considered to be support or oppose the Twitter content and sometimes it appears in the form of a kind of news. However, some news may have no sentiment factors on the surface, but it will suggest readers to make specific choices. Because there is no reliable method to deal with this information, researchers will not consider the neural sentiment. It is important to know that a Twitter may contain many different emotions, the classification should according to the primary emotions which means the emotion that taking account for the largest proportion. In the part, after pre-process of the Twitter content, researchers will compare the Twitter Sentiment and Twitter emotion of the two US presidential candidate Donald Trump and Hillary Clinton respectively. This task can be divided into three parts, first step, we will collect related tweets about each presidential candidate. Then, R programming language will be applied to the Twitter sentiment the result will show in the percentage of positive and negative. At last, lexicon based method will be implemented to classify the Twitter association emotions.

By applying NRC Emotion lexicon in the collected tweets retrieved from Twitter about “Hillary Clinton”, we can obtain the sentiment result (the percentage of positive and negative) and the emotion result (the percentage of different emotion anger, fear, anticipation, trust, surprise, sadness, joy and disgust). Figure 3.7 illustrates the Twitter sentiment about Hillary Clinton. Sample Twitter data analysis about Donald Trump

### Twitter Sentiment Analysis about Hillary Clinton and Donald Trump

In this part, researchers will implement a lexicon of NRC to explore the relationship between Twitter sentiment, Twitter Emotion and the results of the USA presidential election result. More specifically, Twitter often contains a variety of sentiment information, however different sentiment or emotion will lead to a different influence to the election result. We have implemented NRC lexicon to obtain the daily sentiment dataset index and daily emotion dataset from 11/06/2016 to 07/11/2016. According to lexicon, sentiment include Positive, Negative and Neural, emotion include Anger, Anticipation, Disgust, Fear, Joy, Sadness, Surprise and Trust. The sentiment data and emotion data is acquired from the Tweets with the key words of Hillary and Trump. Figure 3.8 and 3.9 below shows the daily sentiment index change of Hillary Clinton and Donald Trump.



Figure 3. 8 Daily sentiment index change of Hillary Clinton



Figure 3. 9 Daily sentiment index change about Donald Trump

As can be seen from the two figures above, Figure 3.8 and Figure 3.9 represent the daily Twitter sentiment data about Hillary Clinton and Donald Trump respectively. The red lines in the two figures represent the variety of the positive Twitter proportion data; the blue lines represent the variety of the negative Twitter proportion data; the green lines represent the variety of the neural Twitter proportion data. For Donald Trump and Hillary Clinton, the popularity statistic is shown in the Table 3.1 below and the opposition statistic is shown in the Table 3.2:

Table 3. 1 Statistics of Trump and Hillary Popularity

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | The Highest | The Lowest | Average | SD |
| Trump | 0.4627(20/07) | 0.2257(10/08) | 0.3182 | 0.042 |
| Hillary | 0.4627(14/07) | 0.1857(03/07) | 0.3154 | 0.0576 |

Table 3. 2 Statistics of Trump and Hillary

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | The Highest | The Lowest | Average | SD |
| Trump | 0.4150(30/09) | 0.1650(20/07) | 0.2909 | 0.0431 |
| Hillary | 0.4680(09/07) | 0.1783(27/07) | 0.2844 | 0.0552 |

Table 3.1 and Table 3.2 describe the Twitter sentiment index of Trump and Hillary. For a more comprehensive understanding of the election status, some important date should be taken into account, such as the date of United States presidential debates (26/09/2017), the date of United States vice presidential debates (04/10/2017), the date of United States presidential debates (second time, 09/10/2017), the date of United States presidential debates (third time, 19/10/2017), the data results of the important date are shown in Table 3.3 below.

Table 3. 3 Popularity of Hillary and Trump on some important dates

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 26/09/2017 | 04/10/2017 | 09/10/2017 | 19/10/2017 |
| Trump | 0.3510/0.2732 | 0.3290/0.3240 | 0.2787/0.3340 | 0.3187/0.2913 |
| Hillary | 0.3020/0.3332 | 0.3270/0.3997 | 0.3250/0.2516 | 0.2817/0.3287 |

According to Table 3.3 above, the first data represents the support ratio and the second data represents the opposition data in each section. After the first and third presidential debate, Donald Trump’s Twitter support ratio is higher than Hillary Clinton, similar like this, Trump’s team lead in the Clinton team’s in Vice Presidential debate. The support ration in Twitter about Clinton’s team is higher than Trump only after the second Presidential debate. Consider about the opposition rate, Clinton only did better than Trump in the second Presidential debate.

In order to explore the popularity trend of these two presidential candidates, researchers have implemented the bar chart to express the difference of the Hillary sentiment indexes and Trump sentiment indexes. The results are shown in Figures 3.10 and 3.11.

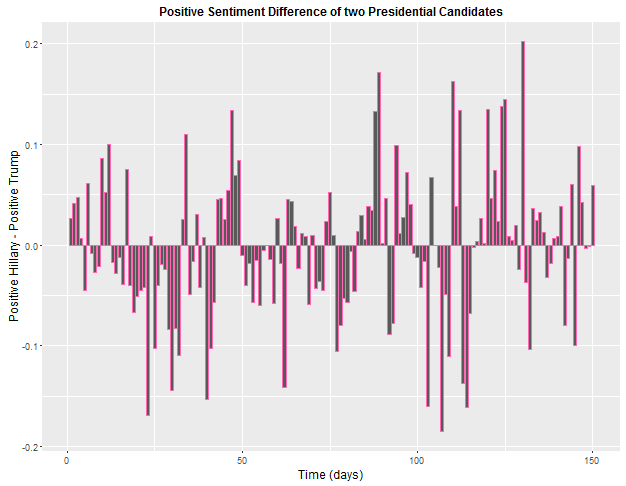


Figure 3. 10 Difference between Clinton and Trump positive Twitter sentiment index

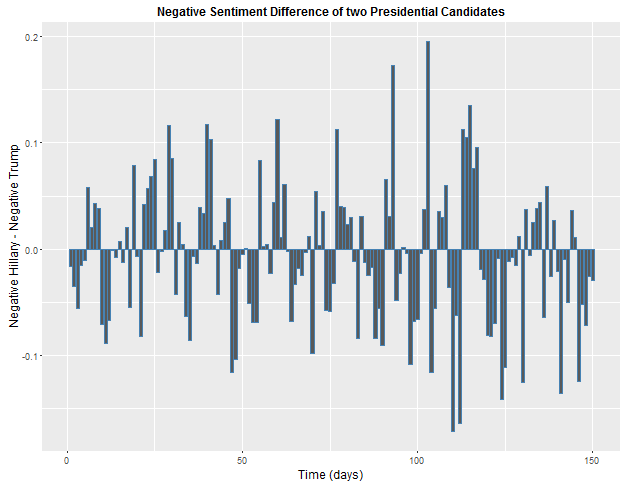


Figure 3. 11 Difference between Clinton and Trump Negative Twitter sentiment index

The above two figures show the Twitter sentiment index difference of Hillary and Trump, how about the emotion index change about Hillary and Trump? In next part, researchers will describe the Twitter emotion index change about Hillary and Trump, researcher use different colour to represent emotion index (Anger, Anticipation, Disgust, Fear, Joy, Sad, Surprise and Trust).

### Twitter Emotion Analysis about Hillary Clinton and Donald Trump

In this part, researchers will implement a NRC lexicon to explore the relationship between Twitter Emotion and the results of the US presidential election result. We have implemented NRC lexicon to obtain the daily emotion dataset from 11/06/2016 to 07/11/2016. According to lexicon, emotion include Anger, Anticipation, Disgust, Fear, Joy, Sadness, Surprise and Trust. The sentiment data and emotion data is acquired from the Tweets with the key words of Clinton and Trump. Figure 3.12 and Figure 3.13 below shows the daily emotion index change of Hillary Clinton and Donald Trump.

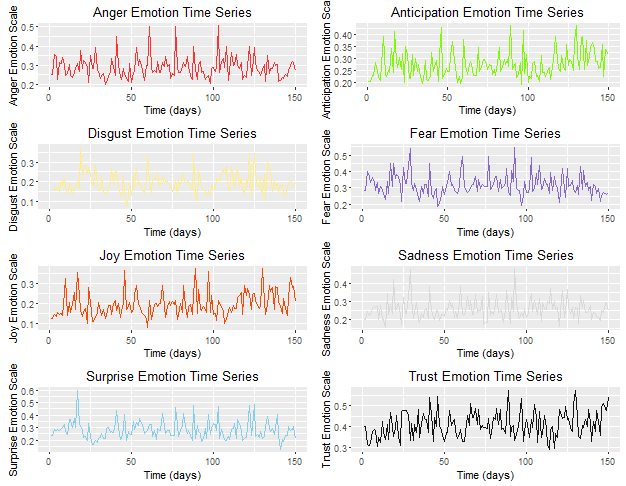


Figure 3. 12 Daily Emotion index about Hillary Clinton

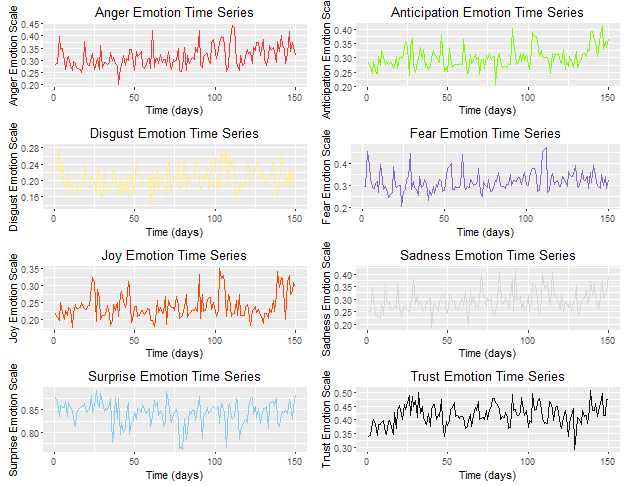


Figure 3. 13 Daily emotion index about Donald Trump

As can be seen from Figure 3.12 and Figure 3.13 above, the different Twitter emotion indexes of Hillary and Trump are most of the emotion index locates in between 0.2 to 0.5 index level, however the Surprise index level shows that Trump’s Surprise index is higher than Clinton’s Surprise index. It means that majority of Twitter user thinks Donald Trump is surprising. Since Donald Trump’s campaign speeches are often surprising, hence our experiment results are consistent with the facts. In order to explore the impact of Twitter emotion to the two presidential candidates. The difference between Trump’s emotion index and Hillary’s emotion index will show in the Figure 3.14 to Figure 3.21 below:

* Clinton and Trump Twitter Anger Emotion Index Difference

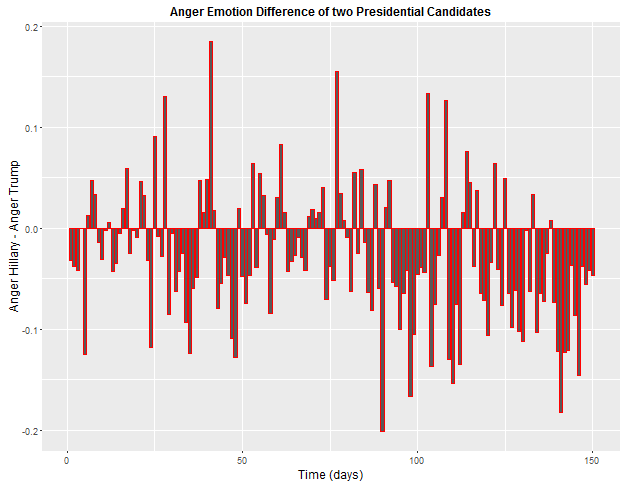


Figure 3. 14 Difference of Twitter anger emotion time series about Hillary and Trump

* Hillary and Trump Twitter Anticipation Emotion Index Difference

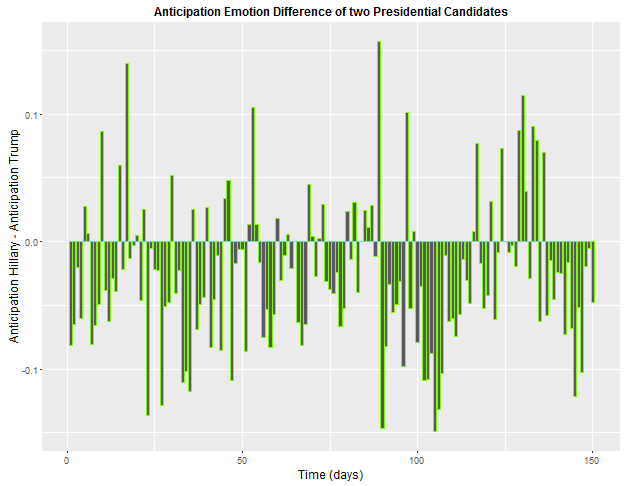


Figure 3. 15 Difference of Twitter anticipation emotion time series about Hillary and Trump

* Hillary and Trump Twitter Disgust Emotion Index Difference

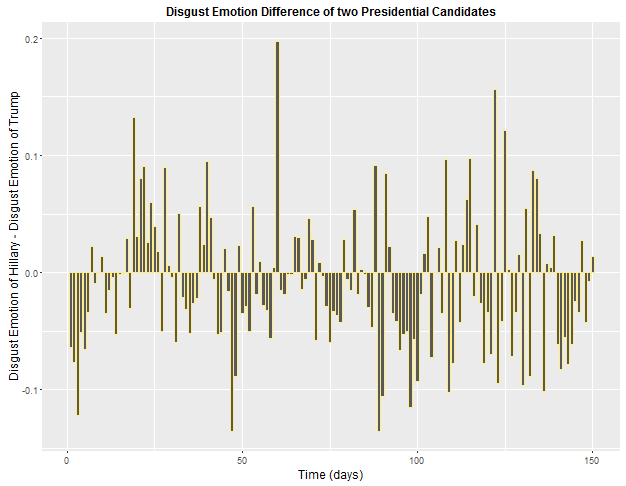


Figure 3. 16 Difference of Twitter disgust emotion time series about Hillary and Trump

* Hillary and Trump Twitter Fear Emotion Index Difference

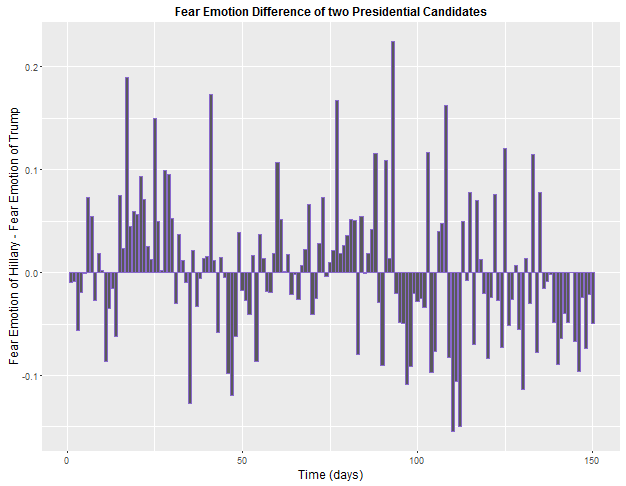


Figure 3. 17 Difference of Twitter fear emotion time series about Hillary and Trump

* Hillary and Trump Twitter Joy Emotion Index Difference

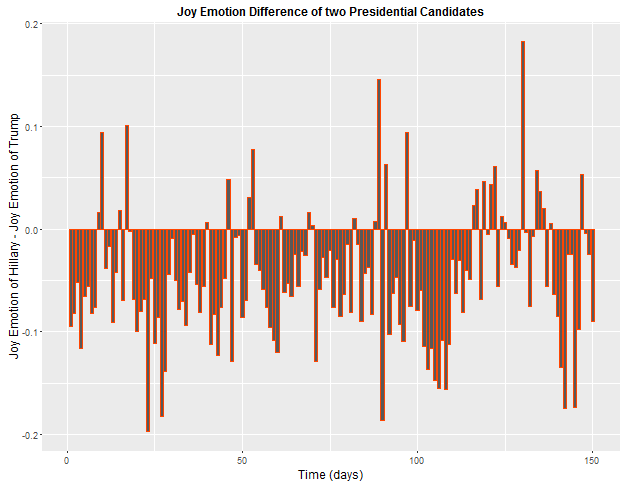


Figure 3. 18 Difference of Twitter joy emotion time series about Hillary and Trump

* Hillary and Trump Twitter Sadness Emotion Index Difference

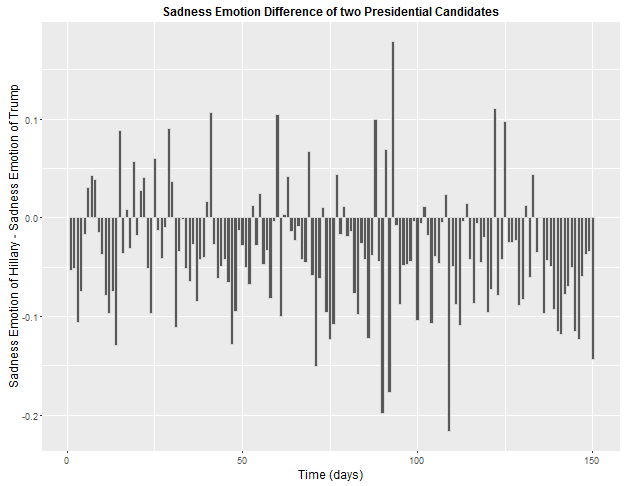


Figure 3. 19 Difference of Twitter sadness emotion time series about Hillary and Trump

* Hillary and Trump Twitter Surprise Emotion Index Difference

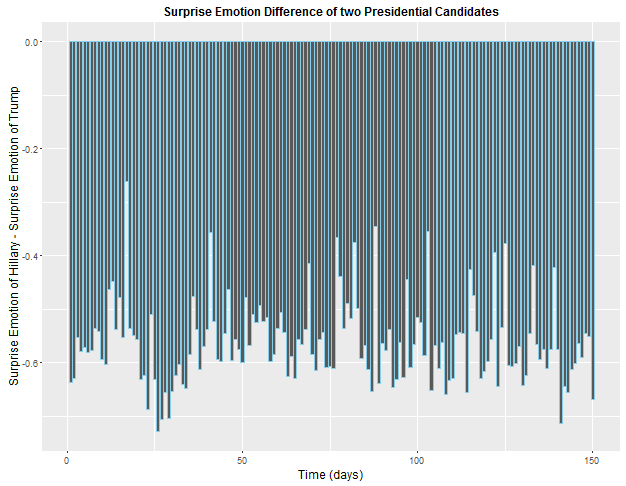


Figure 3. 20 Difference of Twitter surprise emotion time series about Hillary and Trump

* Hillary and Trump Twitter Trust Emotion Index Difference

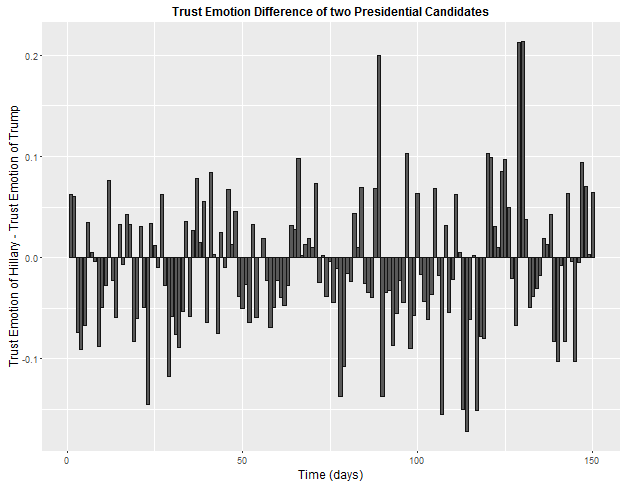


Figure 3. 21 Difference of Twitter trust emotion time series about Hillary and Trump

As can be seen from the figures above, different Twitter emotions about Clinton and Trump have been described. How to describe the Twitter emotions of the two presidential candidates is a challenging task. Researchers prepared to summarize the days of Clinton and Trump in terms of who had a higher emotional index on that day. More specifically, if the difference of Twitter emotion index is positive, it means that Hillary’s Twitter emotion is higher than Trump; on the contrary, if the difference Twitter emotion index is negative, it means that Trump wins on that day. Higher Twitter emotion index is not necessarily a good thing, if a presidential candidate had a higher disgust Twitter index, this means on that day, public opinion showed more disgust emotion about them. In order to have a clear emotional distribution comparison of the two presidential candidates, the researcher used a radar chart. The result is shown in Table 3.4 and Figure 3.22.

Table 3. 4 Twitter emotion distribution by days

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Twitter Emotion | Anger | Anticipation | Disgust | Fear | Joy | Sadness | Surprise | Trust |
| Hillary Clinton | 47 | 42 | 58 | 72 | 29 | 33 | 0 | 64 |
| Donald Trump | 103 | 108 | 92 | 78 | 121 | 117 | 150 | 86 |

Figure 3. 22 Twitter emotion index comparison between Hillary and Trump

As can be seen from Table 3.4 and Figure 3.22, the Twitter emotion distribution by days has been clearly displayed. In Figure 3.22, the blue line represents Hillary while the yellow line represents Trump. Obviously, each Trump’s Twitter emotion are leading Hillary. The results show that there are nearly equal number (Trump 78 and Hillary 72) of fear index; Trump wins on surprise index on every day before the US presidential election. Trump also wins the joy index (121:29) and sadness index (117:33). Through the analysis of the result, Twitter content shows extreme emotion on Trump rather than Hillary which is close to the reality.

## Twitter Sentiment for Brexit 2016

### Introduction

The United Kingdom withdrawal from the European Union (or short for Brexit, UK referendum 2016) has been a significant political topic in the UK and around the world. The UK 2016 referendum had a profound influence on the pattern of the world in economic, political and military. There used to be a heated debate on whether the UK should withdrawal from the European Union or not around the world, especially in the UK. In this part, researchers mainly focus on the Twitter sentiment about this political topic. A total of 23332 tweets about the UK referendum 2016 is collected before 23/06/2016 (The eve of Brexit). Furthermore, the tweets are collected by location (latitude and longitude) and the range. In order to obtain the explicit Twitter sentiment, the population distribution should be discussed. Because of the uneven distribution of population in the UK, it is believed that large and medium cities will have more Twitter users compared to small cities, villages and towns. Figure 3.22 below shows the general UK population distribution. According to the UK population distribution, Tweets about Brexit 2016 are collected by UK regions: London area, Central and North.

### Lexicon based method NRC

1. Twitter Sentiment of Central UK

In this report, the central UK is defined as the middle of the UK that includes big cities such as Leeds, Manchester, Birmingham, Bradford, York, Sheffield, Nottingham and Liverpool. The tweets about the topic Brexit and the geographical coordinates locates in the central UK is extracted, and a total number of 5332 tweets are acquired. Firstly, researchers will consider about the Twitter sentiment which include positive, negative and neural. Figure 3.23 below illustrates the Twitter sentiment about Brexit in the central UK.

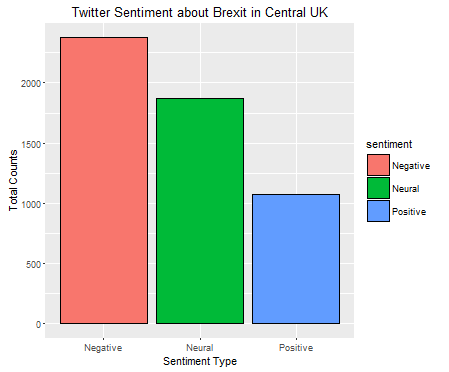


Figure 3. 23 Twitter Sentiment about Brexit in the central UK

As shown in Figure 3.23 above, there are more tweets show negative sentiment about Brexit than positive and there is a large amount of people who show neural sentiment. More specifically, the counts percentage of different sentiment tweets is shown in the table below.

Table 3. 5 Twitter sentiment results in central UK

|  |  |  |  |
| --- | --- | --- | --- |
|  | Positive | Neural | Negative |
| Counts | 1076 | 1867 | 2379 |
| Percentage | 20.18% | 35.02% | 44.62% |

1. Twitter Sentiment of London Area

London is the capital of UK and there are 8.674 million residences living there. A total number of 18000 tweets about Brexit are collected. Implementing the lexicon method in these 18000 tweets, the results of Twitter sentiment of south UK is shown in Figure 3.24.

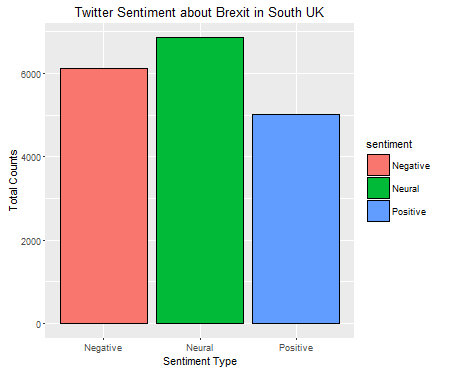


Figure 3. 24 Brexit Twitter Sentiment in London Area

As shown in Figure 3.24 above, more tweets show negative sentiment about Brexit than positive sentiment and there is a large amount of people show neural sentiment. More specifically, the counts percentage of different sentiment tweets is show in Table 3.6 below.

Table 3. 6 Twitter sentiment result in south UK

|  |  |  |  |
| --- | --- | --- | --- |
|  | Positive | Neural | Negative |
| Counts | 5014 | 6860 | 6126 |
| Percentage | 27.86% | 38.11% | 34.03% |

1. Twitter Sentiment of North UK

The north UK is defined as the north part of UK which includes big cities such as Edinburgh, Glasgow, Newcastle. A total number of 6666 tweets are acquired from R. Using lexicon based method to analysis these twitter sentiment, the results north UK twitter sentiment are shown in Figure 3.25 below.

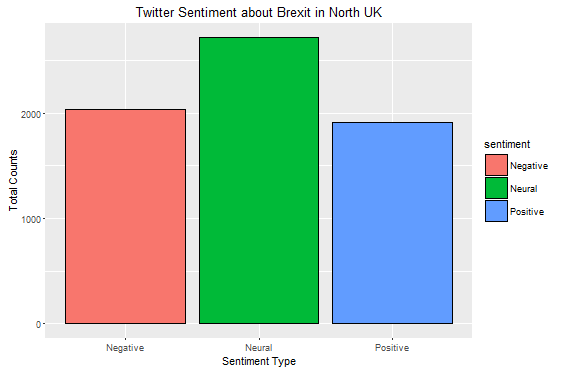


Figure 3. 25 Brexit Twitter sentiment in North UK

According to Figure 3.25 above, although neural Twitter take a large proportion, the positive Twitter is slightly larger than negative Twitter. Except the neural Twitter, the ratio of positive emotion Twitter and negative emotion Twitter are very close to each other. More precise counts percentage of different sentiment tweets is show in the table 3.7 below.

Table 3. 7 Twitter sentiment result in north UK

|  |  |  |  |
| --- | --- | --- | --- |
|  | Positive | Neural | Negative |
| Counts | 1912 | 2722 | 2032 |
| Percentage | 28.7% | 40.8% | 30.05% |

### Results Analysis

In this part, researchers have proposed a method to predict the Twitter public opinion referendum about UK withdrawal from EU by stressing the role of Twitter sentiment in the final decision-making. Traditional way to predict voting result depends on online poll and phone survey which is difficult to acquire comprehensive data. Considering the popularity of Twitter, there is massive data of tweet contains sentiment rich information that have been posted on Twitter. Based on it, this method applies large scale tweet dataset that posted on Twitter by means of lexicon method to measure the sentiment level of the tweets about Brexit and such result is able to help us to predict the referendum results. In order to increase the forecast accuracy, the extracted tweets are all posted in UK. Next, NRC sentiment lexicons are used to distinguish the corpus data into three sentiments (positive, negative and neural). According to the experiment results, there are more tweets show negative sentiment, which means that more people against Brexit on Twitter. However, the support and against rate are very close and a large number of tweets show neural sentiment about Brexit and it will lead to bias for the prediction result. With more reliable method of sentiment analysis, the predictive power will be increased.

## Twitter Sentiment for UK stock market

### Background

In the part, Twitter sentiment and Twitter emotion of FTSE 100 will be visualized and evaluated. Researchers will apply R language to collect tweets that related to FTSE100. Because of the collected tweets include many garbled links and other information that will affect our experiment results, the tweets need to be pre-processed. Then, R programming language will be applied to obtain the FTSE Twitter sentiment. The result will show in the percentage of positive, neural and negative. In addition, the researcher will use a lexicon based method to acquire the tweets’ association emotions (anger, fear, anticipation, trust, surprise, sadness, joy and disgust). Lastly, the result will be visualized and evaluated.

### Data preparation

The experiment FTSE 100 data are chosen from 13/06/2016 to 11/11/2016, considering about the weekends and bank holidays that the stock market will close. There are 110 days’ Twitter data in total of 55666 tweets data for researchers to analysis. The daily tweets are collected by R, and the result will be evaluated in the next part.

### Lexicon based method

Implementing lexicon based method, the FTSE 100 Tweets will be evaluated into sentiment index (include Positive, Negative and Neural) and emotion index (include Anger, Anticipation, Disgust, Fear, Joy, Sadness, Surprise and Trust). Figure 3.25 below shows the daily sentiment index (positive, negative and neural) of FTSE100.

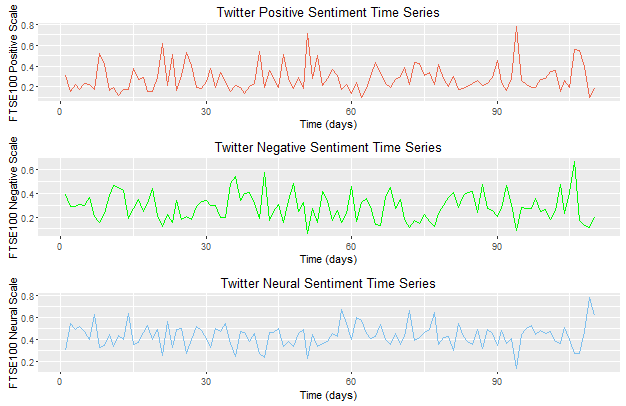


Figure 3. 26 FTSE Twitter sentiment index

As can be seen from Figure 3.26 above, the red line shows the positive Twitter sentiment index change; the green line shows the negative Twitter sentiment index change; the blue line shows the neural sentiment index change.

It is significant to visualize the daily Twitter sentiment situation, we define the difference of Twitter positive index and Twitter negative index as the Twitter polar index which is show in the equation below:

For the purpose of model and predict UK stock market using Twitter sentiment data, the Twitter polar index is important. Researchers will illustrate the Twitter polar index with bar plot which is shown in Figure 3.27 below.

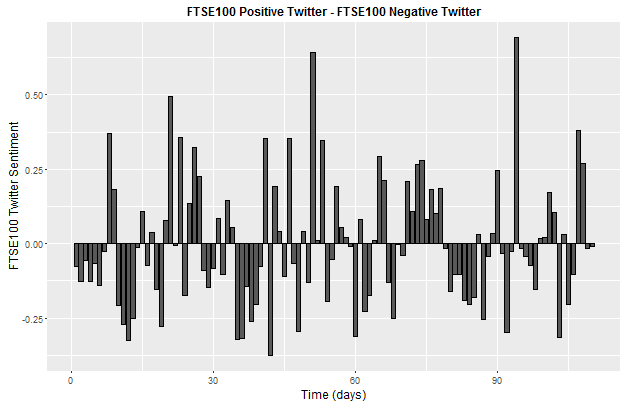


Figure 3. 27 Twitter polar index bar chart

The FTSE Twitter sentiment index is important in modelling and forecasting the FTSE price, however, for an enhanced understanding about the scale of FTSE Twitter data, researchers will explore the Twitter data using emotion lexicon. The Twitter data will be analysed in 8 emotions and the results are show in the Figure 3.28 below:

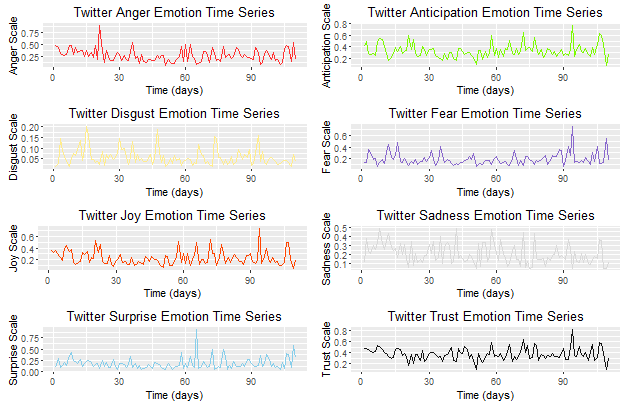


Figure 3. 28 FTSE Twitter Emotion Index

Figure 3.28 illustrates the FTSE Twitter Emotion index change. The subplots show anger index change, anticipation index change, disgust index change, fear index change, joy index change, sadness index change, surprise index change and trust index change respectively. These datasets will be applied to model and forecast the FTSE 100 change in the next chapter.

## Conclusion

In this chapter, researchers explained the process of Twitter data extraction and analysis with R programming. Our Twitter API based in R is able to extract real-time and updated tweets from Twitter. Although the extracted tweets usually include useless or disturbing information such as @ someone, links and different language data, these data can be removed or classification using our text pre-process methods. The application of NRC Lexicon on the semantic analysis of US presidential election Twitter data, Brexit Twitter data and FTSE 100 Twitter provide us important information. We have gained valuable public opinion information for presidential election and UK referendum. The sentiment and emotion index distribution of the two presidential candidates before the election which is proven to be which is really close to reality situation. For example, the daily surprise emotion index of Donald Trump is significantly higher than Hillary Clinton on every single day before the election date; the anticipation and joy daily show Trump wins Hillary on Twitter; Trump’s Twitter also show higher disgust index and anger index and for the fear index, these two presidential candidates present the similar results. In fact, the Twitter opinion results show Trump has the higher emotion index on Twitter than Hillary which means that the Twitter related to Trump gives us more emotions. By summarize the sentiment and emotion index of these two presidential candidates, results show that Trump is more competitiveness on Twitter than Hillary. Although our Twitter model results show that the UK Brexit referendum 2016 is more people support stay in European Union, however, the real referendum results is opposite. The failure in predict the Brexit vote has two reasons: 1. There is not enough Twitter samples for our experiments; 2. The Brexit Twitter has not been extracted day by day, which make us cannot see the changes in public opinions; 3. Not everyone use Twitter to express their opinions. It is believed that by deep mining these Twitter data, we can obtain more information on public opinion. With the help of NRC lexicon, we also get Twitter sentiment indexes and Twitter emotion indexes about FTSE100. For the future research, these opinion-rich datasets can help us to modelling economic problem based on nonlinear models and complex network theory.

# Chapter 4 Machine Learning on Sentiment Analysis and Complex Network

## Introduction and Background

Sentiment analysis also refer to opinion mining and it aims to automatically recognize the sentiments and emotions that contained in the text information. In the previous chapter, researchers focus on Lexicon based method on Twitter sentiment analysis. This chapter will mainly discuss the machine learning methods in Sentiment analysis. Machine learning method is one of the most popular algorithm in text sentiment analysis. The basic idea of machine learning method is using the training data to build a model and the test data will be implemented to the model for classification. Currently, the most popular classification machine is: SVM, Naïve Bayes Classifier, Decision Tree and KNN. The feature selection is the core task of the machine learning algorithm, in this part, a lexicon based feature selection is applied and the performance will be compared and evaluated.

In 4.6 and 4.7, researchers applied complex network analysis on Twitter sentiment data about FTSE 100 close price and R21-15 for data visualization. In this chapter, the tweets that used in previous part will be implemented in complex network analysis, especially in information visualization process. Why these data are significant to visualization process? According to Fekete et.al (2008), information visualization had provided a platform for evaluating quantifiable metrics and these processes could be judged and assessed in a clear and accurate method. Fekete (2008) also stated that there still exist challenges in communicating and recognizing the data. Twitter is a social media network service that used for communication and interaction. Palen and Vieweg (2008) stated that social media interaction is a “highly distributed, decentralized and real time” process. In this chapter, researchers have implemented Complex network analysis to study the Twitter sentiment of UK stock market and discover some interesting features between tweets and authors.

## Twitter Data Pre-process

Before conducting the Sentiment analysis on Twitter, the Twitter data pre-processing is necessary. Because the researchers have already discussed the Twitter data pre-process work in Chapter 3, here it will not be discussed in detail. The processed Twitter data will be directly used in this part.

## Feature Selection for Twitter Data

### Traditional Feature Selection Methods

In the text classification process, text data always have high dimensions. A set of text data can sometimes include thousands of feature vectors and this will affect the classification method. The experiments results show that the classification results will improved with the feature dimensions increase, however, when the feature dimensions are continuing increase, the classification performance will decrease. There is some commonly used text feature selection method such as Document Frequency (DF), Information Gain (IG), Mutual Information (MI). When applying this text feature selection methods, researchers should set a threshold to filtered the inappropriate feature. Here researchers will list some calculation methods of feature weight:

* Term frequency (TF) Weight

TF represents the number of text term’s occurrence in a document. Luhn (1957) states that “The weight of a term that occurs in a document is simply proportional to the term frequency.” The contents, format and length of text are different, all these factors will influence the TF value and the usual method to deal with this problem is normalization. In practical applications, if the text feature includes many stop words (the, an, my…), and the high frequency occurrence of these words will increase the TF weight of stop words, the classification results will be influenced by this. In conclusion, the TF results has strong dependency on removing stop words.

* Inverse Document Frequency (IDF)

Because some English words are commonly used, TF will not fully reflect the text meaning. Sparck (1972) states that “the specificity of a term can be quantified as an inverse function of the number of documents in which it occurs.” Large feature IDF means the feature distribute is very concentrate in the document. In other words, IDF show the quantitative distribution of documents’ feature. The IDF has its rationality, however, the method ignored the dispersion and frequency of the text features.

* Term Frequency – Inverse Document Frequency (TFIDF)

TFIDF describes a numerical series that shows the importance of a term (word) to a document or corpus. Although TFIDF comprehensive considered the advantages and disadvantages of TF and IDF. For instance, “the” is very commonly used in many documents, and the word has a high TF. However, IDF of the word “the” is low. Hence, the considering about TF and IDF the word should be given a low weight. Although TFIDF have many advantages compare to TF and IDF feature extraction, there are still some disadvantages about TFIDF: 1. TFIDF is only effective in lexicon level feature; 2. TFIDF is not able to capture semantics features.

### Feature selection based on NRC lexicon

By applying the NRC lexicon base methods on the Twitter Data, we have found that the NRC lexicon can distinguish the Twitter data into three sentiments (Positive, Negative and Neural) and eight emotions (Anger, Anticipation, Disgust, Fear, Joy, Sad, Surprise and Trust). Because of the machine learning methods will achieve the automatic classification between the polarity of Twitter (Positive or Negative) and this will reflect the public opinion on Twitter, the Twitter emotions are important features for our research. As such, Twitters’ emotional data are our important feature vectors. Each tweet includes more than one emotion in it, as per Figure 4.1 below.

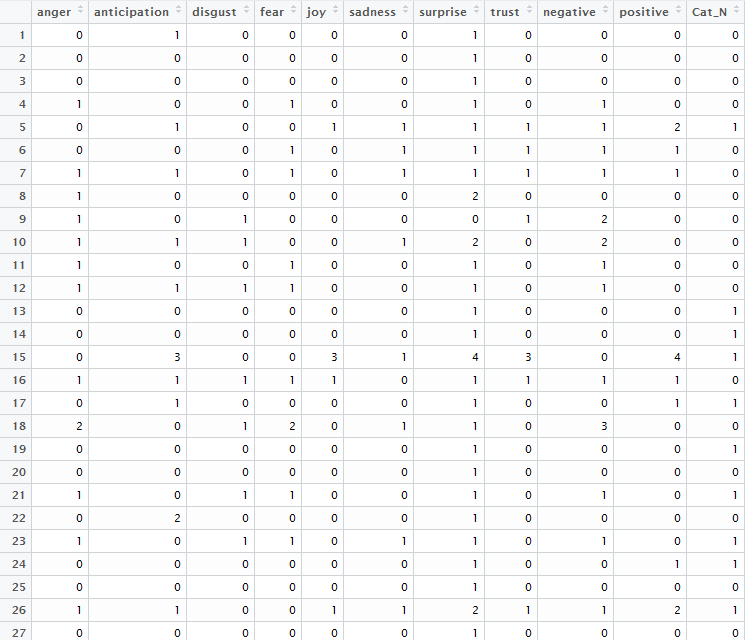


Figure 4. 1 Donald Trump Twitter emotion distribution

As can be seen from Figure 4.1 above, NRC lexicon has helped us to find the emotion features of the Donald Trump’s Twitter. More specifically, each Twitter has been divided into eight emotions. By implementing NRC lexicon, we can obtain the emotion indexes of each tweets and we will use these emotion indexes as our feature for the machine learning methods. The Cat\_N column shows the categories of the tweet: 1 represents positive which means these tweets show approve or support sentiment; 0 represents negative which means these tweets show disagree and against sentiment.

## The Research on Text Classification Algorithm

### Naïve Bayes Classifier

Naïve Bayes (NB) classifier is classification algorithm that based on probabilistic classifier by using Bayes’ theorem. Experience shows that the performance of NB classifier in text classification is better than other machine learning methods, however NB require the independence of the text features. According to Bayes theorem, the NB classifier formula is shown below:

(4.1)

Compared with other machine learning algorithms, the NB is easy to implement with a high classification accuracy and the training time is short. Hence, when the training data increase this algorithm is faster than other algorithms.

### KNN Classifier

KNN and KNN based algorithms for document/text classification have already been widely implemented (Yong, Youwen and Shixiong, 2009) (Trstenjak, Mikac and Donko, 2013) (Bijalwan et al., 2014). The basic principle of KNN is: Assume a sample A text is going to be classified, if there are adjacent k training samples of the A belongs to a category B, then the sample A also belongs to category B. In this algorithm, the chosen k adjacent are already be the correct classification. The k value is significant in the algorithm and need to be given by the algorithm designer. If k = 1, it means the KNN will only chose the nearest neighbour. Low k value will cause disturbance and this will lead to reduction of classification accuracy, meanwhile, high k value will make the classification contains dissimilar samples which will also reduce the classification accuracy. Figure 4.2 below shows the flow chart of KNN classification process:

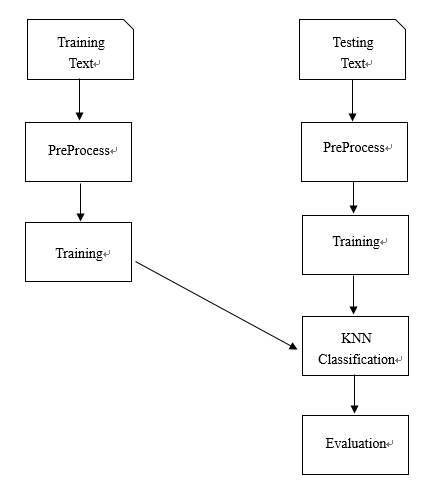


Figure 4. 2 KNN Classification Process

As can be seen from Figure 4.2, the KNN classification is mainly divided into three stages: Pre-Process, Training, Testing. In pre-process, the extracted Twitter data should be arranged tidied, the features are selected by Twitter emotion through lexicon based method. In this part, the Donald Trump Twitter data is chosen as our experiment result. The researcher chose 70% of Donald Trump’s Twitter as the training data and 30% as the testing data. The experiment results are shown in the next part.

## NRC based Machine Learning Methods on Twitter Sentiment Analysis

### Experiment Background

In this part, Because of the time limitation and there is no sample data for our experiments, researchers have manually distinguished 200 tweets that relate to Donald. The experiment data is all from Twitter and it is acquired by Twitter API R. Because of the lack of sample data, researchers cannot compare the experiment results with different training sample, the influence of k value and different feature extraction methods. The 70% of sample tweets data are chosen to be the training data and the 30% of the sample tweets data are chosen to be the testing data. The experiment results are shown in the 4.5.4.

### NRC based KNN Classifier

NRC lexicon on Twitter emotion analysis will provide us eight emotion index results about each Twitter data. More specifically, the NRC lexicon can not only classify the Twitter data into three kinds of sentiments (Positive, Negative and Neural), but also eight emotions (Anger, Anticipation, Disgust, Fear, Joy, Sad, Surprise and Trust). Because of the objective of our KNN classifier is: based on training dataset, KNN will be able to automatically classify the polarity of Twitter (Positive or Negative). The eight Twitter emotions will be used as eight dimensional numeric features for classification. As such, Twitters’ emotional data will be our important feature vectors. The flow chart of our NRC based KNN classifier is shown in figure below:

Twitter Training data

NRC Feature

Extraction

KNN Classifier

Twitter Testing data

NRC Feature Extraction

Evaluation

Figure 4. 3 The process of NRC based KNN classifier

According to the Figure 4.3 above, the first step is choosing the training and texting dataset and the selected texts should include the category label (positive or negative). Then, the feature is decided by NRC emotion index, as well as for other texts. In step three, the K value of KNN algorithm should be determined, the basic rules have already been discussed in 4.4.2, here we do not describe in detail. Next, the classifier determines the category by calculate the nearest kth Euclidean distance between testing data and training data. The Euclidean distance equation is shown below:

(4.2)

Where represents the training points and represents all the testing points, is eight which means eight NRC emotions indexes. The most nearest k point will be selected, and the category will be determined by the category frequency in k.

### NRC based Naïve Bayes (NB) Classifier

Similar like NRC based KNN classifier, NRC based NB classifier will also use NRC emotion features for classification. The process of NB classifier is shown in figure below:

Twitter Training data

NRC Feature

Extraction

NB Classifier

Twitter Testing data

NRC Feature Extraction

Evaluation

Figure 4. 4 The process of NRC based KNN classifier

Naïve Bayes classifier is designed based on statistical theory. In document classification, “the presence or absence of a word in a textual document determines the outcome of the prediction” (Bijalwan et al., 2014). In our experiment, each tweet will be described by a dimensional vector that acquired from NRC emotion lexicon:

(4.3)

The tweets are needed to be classified with a set of (positive and negative) classes:

(4.4)

According to Bayes theory, the probabilities of with a given sample tweet is:

(4.5)

Because of every NRC feature is independent with each other, therefore,

(4.6)

### NRC based KNN and Naïve Bayes Classifier Result Analysis

In order to evaluate the classification results of different machining learning classifiers, researchers will compare common performance index precision, recall, and F-1 score of the model. These values can be acquired from the confusion matrix. **Precision** (also called as Positive Predictive Value) is the fraction of true positive and a predicted conditional positive, **recall** (also known as sensitivity) is the fraction of true positive and condition positive. Additionally, there is a trade-off between Precision and Recall. As a supplement to Precision and Recall, the **F-1** score is implemented.

As we discussed in 2.4.5.7, A represents the condition is positive and algorithm result is positive (true positive); B represents condition is negative while the algorithm result is positive (false positive); C represents the condition is positive while the algorithm result is negative (false negative); D represents the condition is negative and the algorithm result is also negative (true negative). The Precision and Recall are representing in the equation (2.3) and (2.4):

F1 value can be determiner by precision and recall which is shown below:

In the equation above, where represents the Precision, represents the Recall and is the value. We designed two experiments to clarify the NRC based KNN and NB algorithm. The testing and training datasets are chosen randomly and we implement each classifier for ten times. Table 4.1 and 4.2 below illustrates the performance of NRC KNN and NB classifiers:

Table 4. 1 The performance of NRC KNN classifier

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precision | Recall | F1 |
| Experiment 1 | 0.8333 | 0.6452 | 0.7273 |
| Experiment 2 | 0.85 | 0.5 | 0.6296 |
| Experiment 3 | 0.6957 | 0.5517 | 0.6153 |
| Experiment 4 | 0.8077 | 0.6563 | 0.7241 |
| Experiment 5 | 0.64 | 0.5714 | 0.6038 |
| Experiment 6 | 0.9048 | 0.5135 | 0.6552 |
| Experiment 7 | 0.7391 | 0.6538 | 0.6939 |
| Experiment 8 | 0.7222 | 0.5 | 0.5909 |
| Experiment 9 | 0.8182 | 0.5625 | 0.6667 |
| Experiment 10 | 0.84 | 0.5833 | 0.6885 |
| Average | 0.7851 | 0.5738 | 0.6595 |

Table 4. 2 The performance of NRC NB classifier

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precision | Recall | F1 |
| Experiment 1 | 0.7692 | 0.7692 | 0.7692 |
| Experiment 2 | 0.5 | 0.7895 | 0.6122 |
| Experiment 3 | 0.64 | 0.7619 | 0.6957 |
| Experiment 4 | 0.4231 | 0.7333 | 0.5366 |
| Experiment 5 | 0.7083 | 0.6296 | 0.6667 |
| Experiment 6 | 0.75 | 0.6207 | 0.6792 |
| Experiment 7 | 0.4762 | 0.6667 | 0.5556 |
| Experiment 8 | 0.36 | 0.6429 | 0.4615 |
| Experiment 9 | 0.5217 | 0.6667 | 0.5854 |
| Experiment 10 | 0.5833 | 0.6947 | 0.6222 |
| Average | 0.5732 | 0.5738 | 0.6184 |

According to the table 4.1 and 4.2, consider about the average results, the independent experiments illustrates that the NRC based KNN classifier outperform NRC based NB classifier in Precision and F1 value. However, in the experiment 1 of NB classifier, the performance shows the overall best results. The limitation of NRC based classifier is such approaches require large numbers of labelled Twitter data to increase the classification performance. Therefore, when dealing with a novel Twitter sentiment analysis problem, labelled tweets about specific topic is required.

## Twitter Social Network Analysis

### Data Resources

FTSE 100 Tweets are collected by R and the extracted data are stored in excel. For Twitter sentiment, there are some popular and influential tweets that have been applied by other Twitter users to post them on their own Twitter account. Such information is significant because these Tweets is able to influence other Twitter users and public sentiment. As discussed in the previous chapter, R can accurately extract and collect either tweets contents or tweets author data. Furthermore, R can also deal with irrelevant information and rubbish Tweets' data using the Twitter pre-process process. After that, the processed Twitter data will be imported to UCINET to build the data visualization model which include the nodes and links.

### Analysis

Considering about the relationship between sentiment tweets and the authors, the complex networks would be established. Twitter users network about the FTSE 100 in 18/11/2014 is shown in Figure 4.3 below.

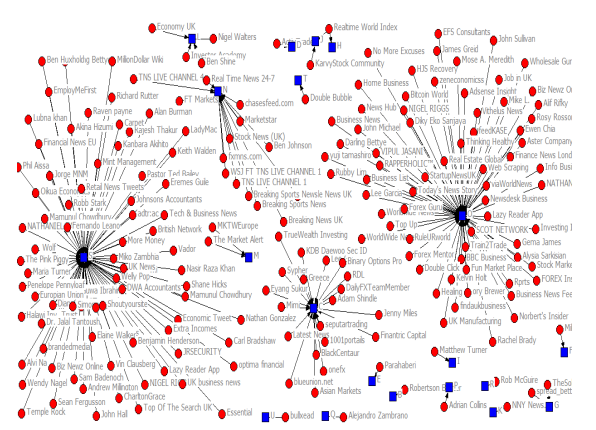


Figure 4. 5 Social network Twitter sentiment about FTSE100 in 18/11/2014

According to Figure 4.3, the blue square nodes are the tweets content and the red circle nodes are the tweets authors. It could be seen from the figure that tweets C, N, S and O are the most popular and influential tweets in 18/11/2014. More specifically, tweet O is “Prudential boosts helps FTSE 100” and there are around 69 users posted this information on their tweets; tweet S is “Energu firms lift FTSE 100” and around 82 users retweet this information. It is clear to see the Figure 4.3 clearly identify the nodes and links about FTSE 100 information in 18/11/2014.

### Summary

The network analysis about FTSE 100 Twitter sentiment in 18/11/2014 show there are some tweets that is more popular than other tweets. These tweets had been retweeted many times by other users. These tweets could actual influence the Twitter sentiment or even public sentiment. Due to the time limitation, more analysis about complex network would be implemented in the future work.

## Conclusion

In this chapter, we simply explored the Machine Learning on sentiment analysis and the data visualization in Complex Network analysis. Considering about the time limitation, researchers cannot study these theories in depth. In the future, when we have enough training sample, the KNN classifier will be applied into US presidential election Tweets, FTSE100 Tweets to obtain a more reliable sentiment index for modelling and predicting. Additionally, a novel and improved KNN classifier is being studied by our group and it is believed that it will have a better performance in the text classification tasks.

# Chapter 5. Stock Market System Modelling – Wavelet Regression Model

## Introduction

In the past few years, stock market research is of great interest and the stock market prediction has been attracting increasing attention from academic and economic. Early studies about stock market prediction are mainly based on random walk theory (Fama, 1965) and news’ information (Qian and Rasheed, 2007). However, the method cannot provide a more than 50% accuracy result (Nofsinger, 2005). It is knowns that the news will affect the stock market change, public opinion also plays an important (Bollen, Mao and Zeng, 2011). According to this, behavioural economics reveals that psychological behaviour plays a significant role in the investment decision making (Marg, 1995) (Dolan, 2002) (Kahneman and Tversky, 2013). When the emotional characteristics appeared in the investors’ decision-making, the public opinion plays an important role in modelling and predicting the stock market change.

According to Tan, Quek and Ng (2005), stock market system is nonlinear, nonparametric, complex and chaotic; Miao, Chen and Zhao (2007) also states that stock market’s varieties are influenced by political issues, economic conditions, bank rate, investors’ sentiment, other stock market price. These stock market system’s features have made it difficult to predict stock market change with traditional nonlinear regression models. As we discussed in Chapter 2, the main feature of the wavelet is stepwise algorithm that can derive a sparse representation of a complex nonlinear system with minimum computation (Billings, 2013). Many properties have made wavelet based regression models are ideal methods for severely nonlinear system identification. This chapter is arranged as follows: Firstly, researchers will apply world stock market systems and wavelet nonlinear models for SSE Composite index system. Secondly, another important economic index crude oil price dataset is used to predict FTSE close price. Lastly, Twitter sentiment and Twitter emotion system will be used as inputs to model FTSE100 close price.

## Shang Hai Composite (SSE) Index Model Representation

In order to explore an algorithm for modelling nonlinear and non-stationary process, Shanghai stock Exchange (SSE) Composite Index is chosen to be the experiment sample. The SEE Composite Index is one of the Chinese stock market index that A shares and B shares are traded in Shanghai Stock Exchange. This index had launched on 15/12/1991 with the base value 100. In this project, researchers choose the SSE Composite index from 04/01/2012 to 31/12/2012 as the output and other historic stock market index SSE, CAC40, DAX, Hang Sheng, SP500 and FTSE100 as the model inputs to establish a Multi Input Single Output (MISO) system. More specifically,

* CAC40 is French stock market index measures the 40 significant values on Paris Bourse;
* DAX is German stock index that build up by 30 German companies;
* Hang Sheng index is another Chinese stock market that traded in Hong Kong;
* SP500 is an American stock market index based on 500 companies;
* FTSE100 is the index of 100 companies of London Stock Exchange;

Modelling and forecasting stock market process is a challenging work, because stock market process tends to be nonlinear, non-stationary, uncertain and it will be influenced by world economic conditions, political policy and investor sentiment. In this study, researchers will only discuss the relationship between world famous stock market price and SSE Composite index. The stock market opening time of SSE and HangSheng are the same whereas the CAC40, DAX, SP500 and FTSE100 are different from SSE. Because we will use linear and nonlinear regression model to predict the stock market change, the input time series HangSheng, CAC40, DAX, SP500 and FTSE100 will be at least one day previous than SSE series, therefore the different time factor of the stock market open time do not need to be considered. In 5.2, some basic wavelet decomposition and wavelet transform will be discussed. In 5.3, researchers will explore the application of Wavelet Multi Input and Single Output model in SSE Composite index process. In 5.4, the model performance will be discussed and evaluated. At last, researchers will implement this model into Twitter - FTSE100 model.

## Wavelet Analysis

### Wavelet background

The significant objective of nonlinear system identification is to obtain an appropriate model based on the input and output variables. This process can be described as implementing polynomial functions, kernel functions and other basis functions with global or local characteristics to construct a nonlinear model. In real world issues, most functions can only be used to approximate certain severe nonlinear behaviour effectively. In some cases, the nonlinearity of the dynamical system cannot be represented at all by a given class of functions because of the lack of good approximation properties. The basis function that is used for approximation should offer some flexibility in adapting the complexity of the model structure, so the model is able to match, as closely as possible, the underlying nonlinearity of dynamic systems.

When the wavelet analysis had been first introduced by Morlet and Grossmann in 1984, it is purposefully created to have the capability that incorporates the global basis function feature and local basis function feature that could be applied in signal processing. Wavelet outperforms Fourier transform and is suitable for arbitrary signals, such as severely nonlinear signals. Fourier transform only explains the frequency domain information and the time information is lost, hence, it is impossible to know that when a specific change of signal take place. Compared with Fourier transform, Wavelet transform has the ability of resolution and localization, which could transform and analyse signals either in frequency domain or time domain, and this could overcome the defect of Fourier transform. Wavelet analysis applies a prototype function, which is called mother wavelet, to decompose a signal into different scales.

### Wavelet transforms

Wavelet transform is able construct a time-frequency representation of a signal that provide good time and frequency localization. Let be a mother wavelet and let the scale and time parameters be represented by and respectively, a continuous wavelet transforms (CWT) is defined as (Mallat, 2008).

(5.1)

Continuous wavelet transform calculates the integration of the product between the original signal and the mother wavelet. The parameter enables the function to shift and locate around . The scale parameter is able to dilate or contract the wavelet function depending on different frequencies. Because of the scale and location information could be acquired from the CWT at other scales and locations, the equation above would lead to signal redundant representation problem. As for the practical application, economic data are discrete signals or time series data rather than continuous signals or time series. In conclusion, discrete wavelet transform (DWT) is often preferred for practical applications. The DWT is shown in equation below (Mallat, 2008)

(5.2)

Discrete wavelet transform is an effective way to avoid signal redundant representation by constraining dilation and location parameters. In the equation above, is a specified dilation parameter which is larger than and is the localization parameter which is positive. The parameters and are all integers that control the dilation and location (Akrami, Mahdi and Santos, 2014). When the parameters meet the condition and , the wavelet is known as dyadic wavelet that is written in the form below,

(5.3)

Let be a time series with period its DWT is a discrete inner product which is shown following,

(5.4)

The discrete wavelet coefficients are the discrete wavelet transform at current scale and location . Thus, with the change of scale and location DWT will provide the variation wavelet coefficients through different scales and locations.

### Selection of Mother Wavelet Function

The selection of mother wavelet function is a research direction in wavelet analysis. However, researchers have not found a well-defined rule that can help us to select a suitable mother wavelet function in a particular application (AI-Qazzaz et al., 2015). In addition, Current studies yet to show a specific mother wavelet function for decomposition of the stock market series (Wadia and Ismail, 2011) (Lee, 2004) (Rua and Nunes, 2009) (Heieh, Hsiao and Yeh, 2011). Despite the lack of reliable rules, the selection of an appropriate mother wavelet usually based on empirical such as wavelet support region, wavelet vanishing moments, similarity and symmetry (Arafat, 2003). Several researches have investigated the Daubechies family wavelets’ application in economic time series analysis: Kao et al., (2013) applied Daubecies 2 (DB2) in feature extraction for stock index; Wadi and Ismail (2011) have implemented DB2 and Haar pre-process the financial time series and they prove that DB2 wavelet gives the best model performance. Therefore, in this project DB2 mother wavelet function will be applied for wavelet decompose stock and other time series.

### Stock Market Data Pre-process Using Discrete Wavelet Transform (DWT)

In this case study, the world wide six stock market indexes, namely, SSE, CAC40, DAX, Hangsheng, SP500 and FTSE100. These large dataset (time series data) will be pre-processed using DWT by the ‘DB2’ mother wavelet at the resolution level of 3. It is emphasized that because of we need to use previous stock market price and previous SSE price. These inputs time series are all at least on day previous of SSE series. The previous inputs will be decided by the model order, the figure 5.1 to 5.6 are all one day ahead wavelet transform. Figure 5.1 illustrates the wavelet transformation of daily FTSE100 index; Figure 5.2 illustrates the wavelet transformation of daily SEE index; Figure 5.3 illustrates the wavelet transformation of daily Hangsheng index; Figure 5.4 shows the wavelet transformation of daily DAX data; Figure 5.5 shows the wavelet transformation of daily CAC data; Figure 5.6 shows the wavelet transformation of daily SP500 data.

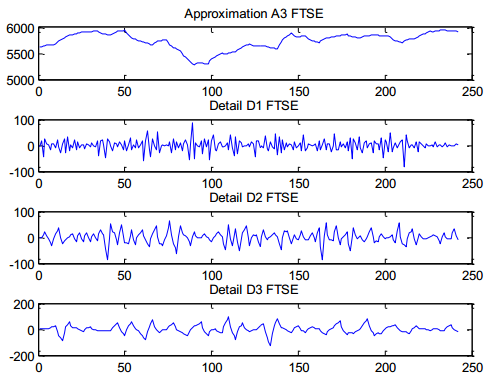


Figure 5. 1 Wavelet Decomposition of FTSE 100 index time series

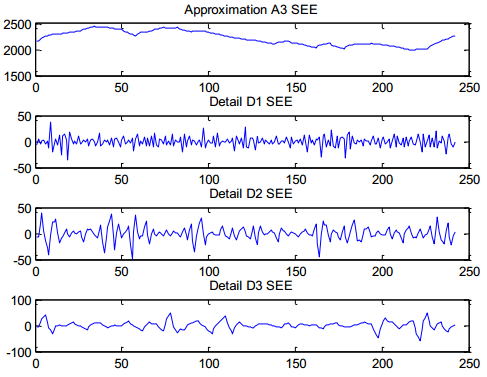


Figure 5. 2 Wavelet Decomposition of SEE Composite index time series

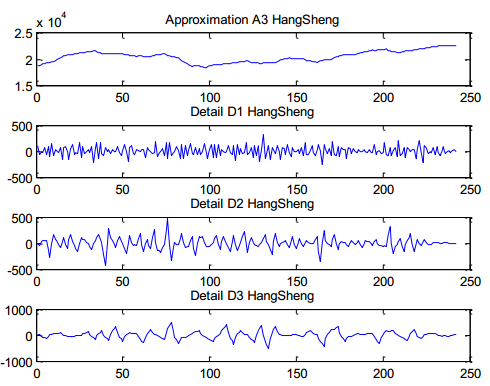


Figure 5. 3 Wavelet Decomposition of HangSheng index time series

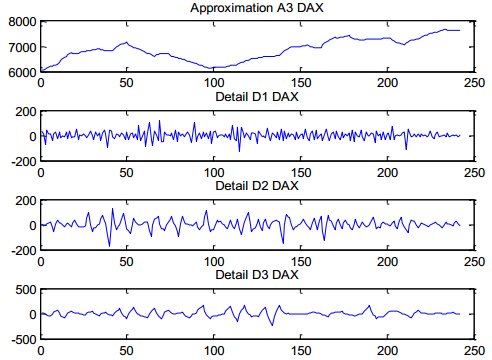


Figure 5. 4 Wavelet Decomposition of DAX index time series

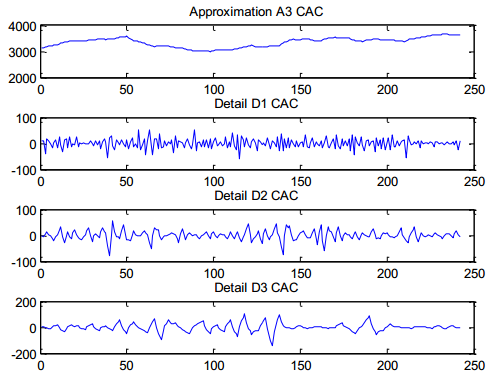


Figure 5. 5 Wavelet Decomposition of CAC index time series

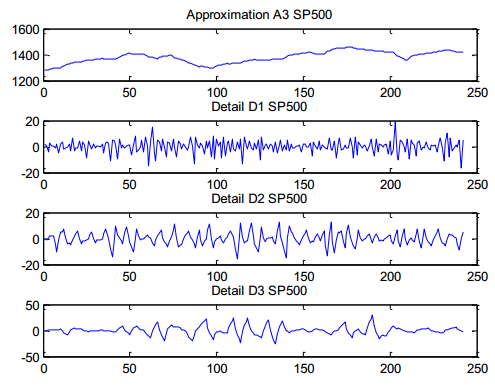


Figure 5. 6 Wavelet Decomposition of SP500 index time series

## Linear Wavelet Multi Input Single Output (WMISO) Model

### WMISO Model Framework

The Wavelet MISO includes Wavelet ARX, Wavelet ARMAX and Wavelet NARMAX models, and these hybrid models are consisted of wavelet part and traditional dynamic regression models (Billings and Wei, 2005). In modelling and predicting tasks, Wavelet MISO method follows the procedures that is shown by figure below. Firstly, all the daily stock market close prices are pre-processed using the wavelet transform. These time series are decomposed into detailed and approximation subseries using “DB2” mother wavelet in the resolution level of three. Secondly, choose the wavelet processed time series as the inputs of the system, then, applying linear and nonlinear methods to model and forecast SSE composite index system at next step.

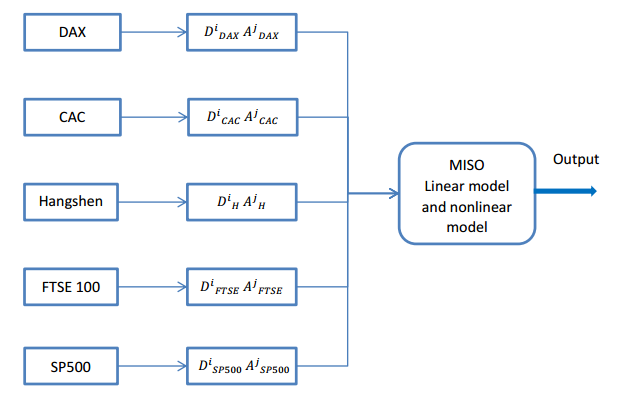


Figure 5. 7 WMISO Model Structure

According to the figure above, represents the detailed subseries at the decomposition leveland represents the approximated subseries the decomposition level .

### Selection of Input Variables

According to the assumption that a given stock market index time series (SEE composite index) is influenced by global economic status. The combination of several economic subsystems will contribute to the SEE composite index. Cross correlation (CC) test is an effective method that usually applied to assess the lag relationship of two variables. Therefore, in this experiment, CC test is used to identify the relationship of lag world stock market close prices series data (daily) and SEE close price series (daily). More specifically, world stock market prices with lag 1 to 5 days and SEE price will be used for CC test and the results are shown in table 5.1 to 5.5. Significant correlation coefficients are identified.

Table 5. 1 Cross correlation analysis about DWT FTSE 100 index and SEE composite index

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 1 Day Lag | 2 Days Lag | 3 Days Lag | 4 Days Lag | 5 Days Lag |
| A3 | -0.2655 | -0.2471 | -0.2310 | -0.2164 | -0.2028 |
| D1 | -0.0064 | 0.0109 | 0.0084 | 0.0181 | -0.0050 |
| D2 | 0.0171 | 0.0434 | 0.0295 | -0.0032 | -0.0320 |
| D3 | 0.035 | 0.0386 | 0.0340 | -0.0041 | 0.0087 |

Table 5. 2 Cross correlation analysis about DWT HangSheng index and SEE composite index

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 1 Day Lag | 2 Days Lag | 3 Days Lag | 4 Days Lag | 5 Days Lag |
| A3 | -0.2132 | -0.1976 | -0.1852 | -0.1727 | -0.1582 |
| D1 | -0.0136 | 0.0581 | -0.0345 | 0.0106 | -0.0089 |
| D2 | 0.0142 | 0.0420 | 0.0290 | -0.0004 | -0.0185 |
| D3 | 0.0660 | 0.0455 | 0.0408 | -0.0040 | 0.0161 |

Table 5. 3 Cross correlation analysis about DWT DAX index and SEE composite index

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 1 Day Lag | 2 Days Lag | 3 Days Lag | 4 Days Lag | 5 Days Lag |
| A3 | -0.5651 | -0.5583 | -0.5538 | -0.5500 | -0.5423 |
| D1 | -0.0080 | 0.0070 | -0.0067 | 0.0262 | -0.0159 |
| D2 | 0.0138 | 0.0341 | 0.0278 | -0.0018 | -0.0200 |
| D3 | 0.0275 | 0.0342 | 0.0368 | -0.0020 | -0.0023 |

Table 5. 4 Cross correlation analysis about DWT CAC index and SEE composite index

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 1 Day Lag | 2 Days Lag | 3 Days Lag | 4 Days Lag | 5 Days Lag |
| A3 | -0.4384 | -0.4278 | -0.4179 | -0.4107 | -0.4013 |
| D1 | -0.0131 | 0.0061 | 0.0103 | 0.0006 | -0.0017 |
| D2 | 0.0095 | 0.0359 | 0.0292 | -0.0034 | -0.0225 |
| D3 | 0.0149 | 0.0248 | 0.0320 | -0.0276 | -0.0272 |

Table 5. 5 Cross correlation analysis about SP500 index and SEE composite index

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 1 Day Lag | 2 Days Lag | 3 Days Lag | 4 Days Lag | 5 Days Lag |
| A3 | -0.5437 | -0.5427 | -0.5412 | -0.5409 | -0.5359 |
| D1 | 0.0016 | 0.0399 | 0.0097 | -0.0097 | -0.0024 |
| D2 | 0.0132 | 0.0368 | 0.0196 | -0.0110 | -0.0216 |
| D3 | 0.0292 | -0.0002 | 0.0440 | -0.0299 | 0.0165 |

In input time series pre-processing step, the wavelet decompositions of 6 world popular stock market series are performed in figure 5.1 to 5.6. The figures clearly illustrate how the original series are decomposed into approximation series and detail series by wavelet. After this, CC test is used to investigate the relationship between these wavelet decomposed lag subseries and SEE close price. Table 5.1 to 5.2 illustrates the summary of correlation coefficient. Compared with other subseries, the DAX index components approximation with lag of 1 day (A1), approximation with lag of 2 days (A2), A3, A4 and A5; The CAC index components A1, A2, A3, A4 and A5; SP500 index components A1, A2, A3, A4 and A5 are having cross correlation with SEE index.

### Wavelet ARX and Wavelet ARMAX

In this section, linear system identification model ARX (autoregressive with exogenous input) and ARMAX (autoregressive moving average with exogenous input) will be applied and combined with DWT to produce our WARX and WARMAX models. The input variables are chosen by cross correlation test between the wavelet decomposition of the influential stock market index and the SEE composite index. Figure 5.8 shows the detailed structure of WARX and WARMAX model.

### Model Structure and Results Analysis

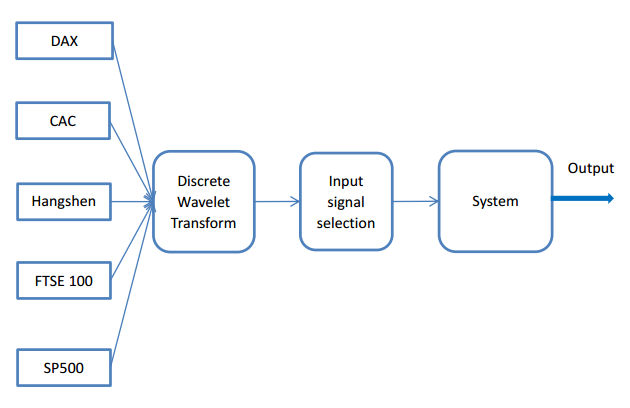


Figure 5. 8 Wavelet linear regression model framework

Because there are twelve inputs variables are selected for WMISO system, the specific models of WARX and WARMAX are accordingly written to be:

(5.5)

(5.6)

The six stock market indices in 2012 are used to implement out experiment. Considering about the holidays and stop plate date, there are 242 data for each stock market close price. The first 200 data are chosen to be the training and the last 40 is chosen to be the validation data. Akaike’s Information Criterion (AIC) is a method that is able to measure the model quality based on different data set and the most accurate model will provide lower AIC value. Generally, choosing the model orders is trade off between model complexity and model performance. By AIC method, the model orders for WARX are chosen , and and orders for WARMAX are chosen , and . The performance of the WARX and WARMAX models for the SEE composite indices is shown in the Figure 5.9 and 5.10.

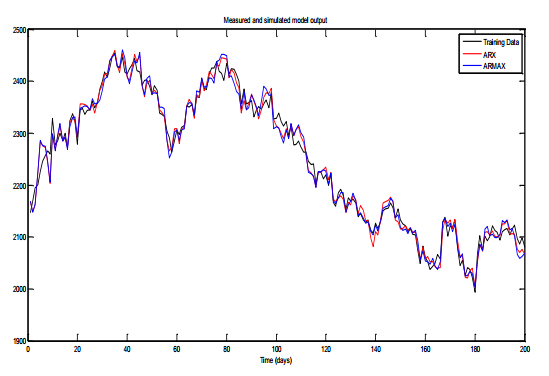


Figure 5. 9 WARX and WARMAX training model result

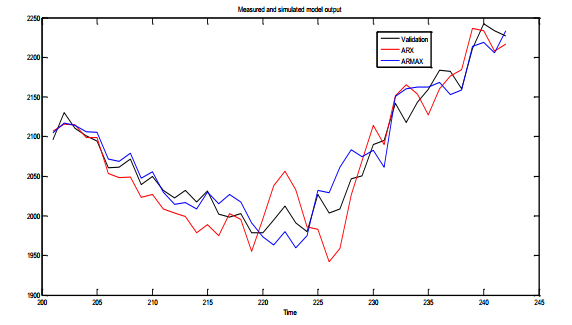


Figure 5. 10 WARX and WARMAX validation model result

Figure 5.9 and Figure 5.10 show simulation result of the one-day ahead predictions from WARX and WARMAX. The black solid line represents the validation (observed) data, the red line is the WARX model output and the blue solid line is the WARMAX model output. The value of two error measurements, namely, mean absolute error and root mean square error for WARX and WARMAX are shown in the table below.

Table 5. 6 One day ahead prediction of WARX and WARMAX model on SSE composite index

|  |  |  |
| --- | --- | --- |
|  | WARX | WARMAX |
| MAE | 19.40 | 17.77 |
| RMSE | 37.53 | 22.20 |

In this part, wavelet based linear models ARX and ARMAX model have been explored and compared, the results illustrate that the Wavelet ARMAX model have improve the MAE by nearly 10% and RMSE by nearly 70%. In 5.5, Wavelet based Nonlinear ARX model will be investigated and the model performance will be compared with wavelet linear models.

## Nonlinear Wavelet Model

The decomposed stock market time series and historical SSE index are used to construct the system input variables. In this study, we use “DB2” as the mother wavelet at 3 decomposition levels. The specific model structure is shown in the Figure 5.11 below.

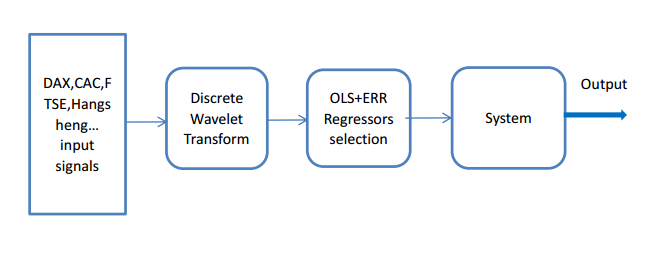


Figure 5. 11 Nonlinear Wavelet Model Structure

The individual input signals have been decomposed by wavelet to produce a new system input. The decomposed signals can be regarded as the multi input time series of the system. Ignore the noise model, the new input signals can be modelled by multi input signal output (MISO) NARX system.The initial nonlinear full model may involve a great number of candidate model terms, but not all the candidate model terms are equally important in representing the system output. Therefore, the Orthogonal Least Square (OLS) method is implemented and used for model refinement by finding important regressors. Consider a nonlinear autoregressive (NARX) model which is shown below:

(5.7)

Equation 5.7 describe the nonlinear ARX model, where and are the system inputs, output and noise. This NARX about SEE system implies that the current output of SEE price () is predicted by its past output values and past input values . describe the previous input terms that are applied to predict current output, describe the previous terms that are applied to predict current output. In many cases, the nonlinear model can be represented by linear-in-the-parameters form:

(5.8)

are the model terms and are the model parameters.

### Orthogonal Least Square Method

The orthogonal least square method had been first developed by Billings and co-workers in the late 1980s and the OLS method is used for parameter estimation of nonlinear models. The basic idea of OLS method is selecting the appropriate inputs that have the maximum influence on the system output. The basic concepts of OLS algorithm is shown in detail below.

Consider a linear-in-the-parameter model in equation below

(5.9)

Where is the output, is the model input, is the noise term and is the model parameters which are going to be estimated. Assume there are outputs , researchers are able to get the form of the linear model, which is shown below.

(5.10)

Where is the output, is the model input, is the noise term and is the model parameters that is needed to be estimated. Assume there are outputs researchers can get from the linear model,

(5.11)

Or in matrix form

(5.12)

Then, we transform into orthogonal vectors , and can be expressed by ,

(5.13)

Because of orthogonal basis spans the same space as the basic set , then can be expressed as:

(5.14)

Considering are orthogonal to noise. The output variance/energy can be written as:

(5.15)

Obviously that the energy of is described by and parts. Because of the noise part cannot be explained, we can only use the ratio which is error reduction ratio that is shown in the equation below:

(5.16)

According to Billings (2013), the ERR “provide a very simple but effective means of determining a subset of significant regressors and the significant terms can be selected according to the value of ERR”. The search will be stopped when the error signal ratio (ESR) is smaller than the threshold that defined at first.

### Model Validation

In this part, researchers will propose a wavelet nonlinear model for SSE stock market system, the predictive will be compared to the wavelet ARX and ARMAX model. The six worldwide stock market indexes are applied in this experiment. Considering the holidays and stop plate data of each stock market, choosing 242 open plate data for each stock market. The first 200 are chosen to be the training and the last 40 is chosen to be the validation data. In MISO NARX model, model terms are selected by OLS + ERR method, large model orders and always lead to large numbers of regressor terms being evaluated. Therefore, in this experiment, the model order is chosen to be and . Setting the threshold is , the result of the OLS + ERR and model terms are shown in table 5.7

Table 5. 7 Identification of SEE system

|  |  |  |  |
| --- | --- | --- | --- |
| Index | Model Terms | Parameter | ERR |
| 1 | SEE(t-1) | 1.4031 | 99.54 |
| 2 | SEE(t-1)SEE(t-3) | -3.05e-4 | 0.15 |
| 3 | A3\_DAX(t-1) | -1.1014 | 0.091 |
| 4 | SEE(t-3)D3\_FTSE(t-1) | 2.94e-4 | 0.075 |
| 5 | SEE(t-2)SEE(t-2) | 3.6932 | 0.012 |
| 6 | SEE(t-2)A3\_SP500(t-1) | -0.0018 | 0.0032 |
| 7 | SEE(t-3)D3\_DAX(t-1) | 8.17e-5 | 0.0083 |
| 8 | SEE(t-1)D2\_HS(t-1) | 0.0027 | 0.0016 |
| 9 | A3\_CAC(t-1) | 5.5637 | 0.0014 |
| 10 | A3\_HS(t-1) | -0.5222 | 0.0008 |
| 11 | SEE(t-2)D2\_HS(t-1) | -0.0048 | 0.0007 |
| 12 | D3\_CAC(t-1) | -5.3398 | 0.0003 |
| 13 | D3\_FTSE(t-1) | 5.111 | 0.0005 |
| 14 | A3\_FTSE(t-1) | -0.9781 | 0.0002 |
| 15 | D2\_FTSE(t-1) | 1.5097 | 0.0002 |

The simulation result for training and validation data are shown in Figure 5.12 and 5.13,

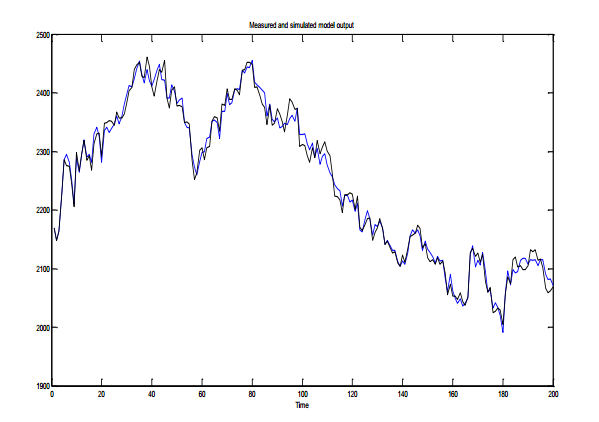


Figure 5. 12 Simulation result of training data

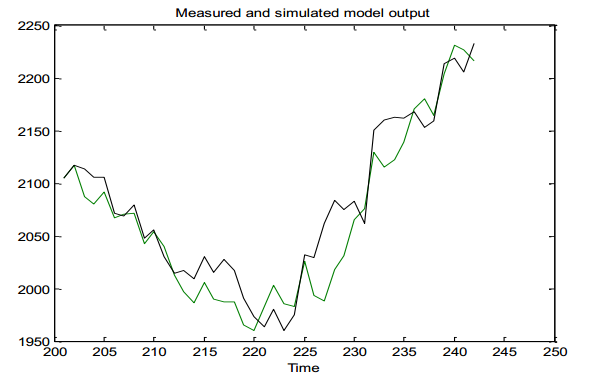


Figure 5. 13 Simulation results of validation data

Figure 5.13 describes the 1 step ahead prediction from the WNARX model. The green line is the prediction result and the black line is the observed data. The root mean square error and mean absolute error of the wavelet nonlinear model is 21.56 and 17.20. The statistic predictive power of wavelet nonlinear model, wavelet linear models and linear regression models are shown in table 5.8 for SSE composite close price from 01/01/2012 to 31/12/2012. Compared with the linear wavelet based method, nonlinear wavelet method has slightly improved the predictive accuracy by reducing the MAE and RMSE.

Table 5. 8 Model performance for SEE system

|  |  |  |  |
| --- | --- | --- | --- |
|  | WARX | WARMAX | WNARX |
| MAE | 19.40 | 17.77 | 17.20 |
| RMSE | 37.53 | 22.20 | 21.56 |

In 5.5, we have highlighted wavelet based nonlinear ARX model, Orthogonal Least Square and error reduction ration have been used to choose the most significant terms for our nonlinear model. The model performance will be evaluated in mean absolute error and root mean square error. Table 5.8 tell us that wavelet based NARX model will provide the best modeling results compared with wavelet based ARX and ARMAX. More specifically, WNARX has significantly decrease the MAE, RMSE of WARX by 12% and 42.55. WNARX has slightly decrease the MAS, RMSE of WARMAX by 3% and 2.8%. Results show Wavelet based nonlinear model can be used in modeling severely nonlinear and un-stationary stock market system.

## Crude Oil price & FTSE100 Wavelet Model

1. Background and Introduction

Crude oil price is a key factor that has significant impact on world economic situation. Stock market price is a primary index to measure current economic condition of a country or a region. Instead of using nonlinear System Identification method and Artificial Neural Network (ANN), this paper will implement linear and nonlinear wavelet models that use oil price index as a system input to predict daily stock market price. The algorithm includes Discrete Wavelet Transform (DWT) and System Identification theory (ARX, ARMAX, NARX and NARMAX). The model performance will be measured by root mean square error (RMSE) and mean absolute error (MAE). There are two major findings of our research: First, we find that oil price can help is to model stock market. Second, wavelet models are proven to be more effective than traditional System Identification model in stock market system.

1. Data Preparation

In this paper, author will analyse weekly and daily relationship between crude oil price and FTSE 100 index applying wavelet nonlinear models. More specifically, the daily datasets are chosen from 29/04/2014 to 12/06/2015; the weekly datasets are chosen from 04/01/2010 to 08/06/2015. Considering weekends, holidays and bank holidays, there are 284 datasets for either daily data or weekly data.

1. Model Structure

Due to the review of multiple linear and nonlinear models, we will apply a hybrid wavelet nonlinear model for this research. Figure 5.14 will illustrate the detailed structure of the wavelet hybrid model.

FTSE(t-1)

FTSE(t-2)

FTSE(t-3)

FTSE(t-4)

FTSE(t-5)

OP(t-1)

OP(t-2)

OP(t-3)

OP(t-4)

OP(t-5)

Discrete Wavelet Transform

(DWT)

OLS + EER

Regressors

Selection

System

Output

Figure 5. 14 Nonlinear wavelet model structure

As we can see from the figure 5.14, author choose the FTSE 100 index (FTSE) and Crude Oil price (OP) with lag 1 to lag 5 as the system input variables. Then, these input variables will be decomposed by discrete wavelet transform (DWT) with ‘DB3’ mother wavelet at 4 decomposition level. The individual input signals (FTSE100 and Oil Price) have been decomposed to detailed time series and approximation time series. These new system inputs can be regarded as the multi input time series of the system. Therefore, the nonlinear model of FTSE and OP system can be expressed in the function below,

(5.17)

In the equation above, is the time series of FTSE100 stock market price. means the detail time series of input , means the approximation time series of input which describes the decomposition depth varies from to and is the decomposition level. In order to choose the most significant terms of stock market system, orthogonal least square and error reduction ration are applied which will be discussed in the next section.

1. Orthogonal Least Square and Error Reduction Ratio

Orthogonal Least Square (OLS) and Error Reduction Ratio was first introduced by Billings and had been used for selecting and estimating significant regressor terms and corresponding kernels of nonlinear models. The basic idea of OLS is choosing the appropriate system inputs that have the maximum influence to the system output. After wavelet decomposition, there are 50 subseries. Based on cross correlation test, there are 6 input variables are chosen from 50 subseries. And the 6 input variables will be regressed using 2nd order NARX model, it leads to an estimation of 720 regressor terms in total.

1. Model output

In this section, we will propose wavelet nonlinear model output for daily FTSE100 and OP system and weekly FTSE100 and OP system, the result will be compared and evaluated. As we discussed in section 2, the daily data of FTSE and OP are chosen from 29/04/2014 to 12/06/2015 and the weekly data are chosen from 04/01/2010 to 08/06/2015. Consider about the holidays and bank holidays, there are 279 data for either daily data or weekly data. Let the first 210 data to be the training data and the last 69 data to be the validation data. In MISO NARX model, the regressor terms are selected by OLS + ERR method. Set the threshold to be 0.01, the result of daily and weekly FTSE & OP model terms and parameters are shown in table 5.9 and 5.10.

Table 5. 9 Identification of Daily FTSE OP system

|  |  |  |  |
| --- | --- | --- | --- |
| Index | Model term | Parameter | ERR |
| 1 | D1\_FTSE(t-2) | 1.0664 | 0.95 |
| 2 | D1\_FTSE(t-3) | -0.8705 | 0.03 |
| 3 | D1\_FTSE(t-1) | 0.9917 | 0.01 |
| 4 | D2\_FTSE(t-2) | 0.8040 | 0.005 |

Table 5. 10 Identification of weekly FTSE OP system

|  |  |  |  |
| --- | --- | --- | --- |
| Index | Model term | Parameter | ERR |
| 1 | D2\_FTSE(t-2) | 1.0011 | 0.93 |
| 2 | D2\_FTSE(t-1) | 1.0077 | 0.06 |
| 3 | D1\_FTSE(t-1) | 1.0029 | 0.005 |

The daily and weekly simulation result for validation is show in figure 5.15 and 5.16.



Figure 5. 15 Simulation results of daily FTSE & OP model

Figure 5.15 describes the 1 step ahead prediction of daily FTSE&OP system based on WNARX model. The blue line is the prediction output and the black line is the observed data. The predictive power is measured by root mean square error (RMSE) and mean absolute error (MAE). Therefore, the RMSE and MAE of daily FTSE&OP model is 19.5893 and 13.8155.



Figure 5. 16 Simulation results of weekly FTSE & OP model validation

Figure 5.16 describes the 1 step ahead prediction of weekly FTSE&OP system based on WNARX model. The blue line is the prediction output and the black line is the observed data. The RMSE and MAE of weekly FTSE&OP model is 55.7490 and 43.7483.

1. Conclusion

In this part, a wavelet nonlinear model is applied and discussed in modelling daily and weekly FTSE100 close price variation. The wavelet nonlinear model is consisted of two parts: first part is discrete wavelet transform which using appropriate mother wavelet to decompose input variables; second model validation which use 2nd order MISO NARX model to model the FTSE&OP system. Choosing the same scale of weekly and daily datasets. Two system output are compared and evaluated. The result show that the predictive power of daily model is significant better than weekly model. More specifically, the daily model had decreased the MAE by 64.86% and reduced the RMSE by 68.42%. The result show that FTSE&OP system performed better in short term forecast than long term.

## Twitter Sentiment and Twitter Emotion Predict Stock Market

1. Wavelet decomposition of Twitter Sentiment and Twitter Emotion

There are many journals and articles show that Twitter can help to predict stock market change. In this part, Twitter Sentiment index (Positive and Negative) and Twitter Emotion index (Anger, Anticipation, Disgust, Fear, Joy, Sadness, Surprise and Trust) that we get from Chapter 3 will be decomposed by Discrete Wavelet Transform (DWT) using ‘DB2’ mother wavelet at 3 resolution level. It is emphasized that because of we need to use previous Twitter sentiment/emotion index price and previous FTSE100 price. These inputs time series are all at least on day previous of FTSE100 series. The previous inputs will be decided by our nonlinear model order, the figure 5.17 to 5.26 are all one day ahead wavelet transform. Figure 5.17 to Figure 5.26 show the wavelet decomposition of Twitter Sentiment index and Twitter Emotion index respectively.



Figure 5. 17 Wavelet Decomposition of Twitter positive index



Figure 5. 18 Wavelet Decomposition of Twitter negative index



Figure 5. 19 Wavelet Decomposition of Twitter anger index



Figure 5. 20 Wavelet Decomposition of Twitter anticipation index



Figure 5. 21 Wavelet Decomposition of Twitter disgust index



Figure 5. 22 Wavelet Decomposition of Twitter fear index



Figure 5. 23 Wavelet Decomposition of Twitter Joy index



Figure 5. 24 Wavelet Decomposition of Twitter sadness index



Figure 5. 25 Wavelet Decomposition of Twitter surprise index



Figure 5. 26 Wavelet Decomposition of Twitter trust index

1. Wavelet Twitter FTSE Model Structure

In this case study, researchers will implement decomposed FTSE Twitter sentiment and FTSE historical data, FTSE Twitter emotion and FTSE historical data to model and predict FTSE close price change. Clearly, these two systems are wavelet MISO model. More specifically, the Twitter sentiment and Twitter emotion will be decomposed into detailed subseries and approximation subseries using ‘DB2’ mother wavelet with the resolution level of 3. Then, researchers will use nonlinear ARX method to model FTSE price change. The sentiment FTSE model and emotion FTSE model are shown in the figures below.

Positive

FTSE previous

Negative

Inputs

Nonlinear ARX

Figure 5. 27 Wavelet nonlinear Twitter Emotion FTSE model structure

Anger

Anticipation

Disgust

Joy

Sadness

Surprise

Trust

Fear

FTSE Previous

Inputs

Nonlinear ARX

Figure 5. 28 Wavelet nonlinear Twitter Sentiment FTSE model structure

1. Data Modelling and Prediction

In this part, researchers will implement FTSE sentiment data, FTSE emotion data and FTSE historical data to model our system respectively. The result will be compared and evaluated separately. The training FTSE and Twitter sentiment data are chosen from 13/06/2016 to 23/01/2017. After removing holidays and bank holidays, there are 158 data for daily FTSE close price and Twitter sentiment index. Researchers choose the first 135 data for training data and the last 23 data for evaluating.

1. Twitter Sentiment & FTSE model results

As we discussed before, Twitter sentiment include the positive sentiment and negative sentiment. In 5.6.4, researchers have already decomposed the Twitter sentiment data and FTSE historical data into approximation level and decomposition levels. By applying these datasets for our system model, the results are shown below:



Figure 5. 29 Simulation results of daily FTSE & Twitter sentiment model



Figure 5. 30 Simulation results of FTSE & Twitter sentiment model validation

Implementing our wavelet nonlinear model, the regressor terms are selected by OLS + ERR method. Set the threshold to be 0.001, the selection result of Twitter sentiment model is: A3\_FTSE, D3\_FTSE, D2\_FTSE, D2\_Positive, D3\_Positive and D1\_FTSE. The simulation results of training and testing are shown in figure above. As we can see from Figure 5.29, the blue line describes the FTSE 100 close price change and the red line is the one step ahead prediction of our training set. Figure 5.30 describe the model output and FTSE 100 variety of texting data. Similar like training model, the blue line describes the FTSE data and the red line is the wavelet model output. The result of the wavelet model performance is: the root mean square error is 14.0519 and the mean absolute error is 11.1159 for training model.

1. Twitter Emotion and FTSE model

As we talked before, Twitter emotion data include: anger index, anticipation index, disgust index, fear index, joy index, sadness index, surprise index and trust index. In this part, researchers will apply wavelet decomposed Twitter emotion data and FTSE historical data for OLS algorithm and the results are shown below:



Figure 5. 31 Simulation results of daily FTSE & Twitter emotion model



Figure 5. 32 Simulation results of FTSE & Twitter sentiment model validation

Implementing our wavelet nonlinear model, the regressor terms are selected by OLS + ERR method. Set the threshold to be 0.001, the selection result of Twitter emotion model is: A3\_FTSE, D1\_FTSE, A3\_Disgust, A3\_Sadness, D3\_FTSE, D2\_FTSE. The simulation results of training and testing are shown in figures above. As we can see from Figure 5.31, the blue line describes the FTSE 100 close price change and the red line is the one step ahead prediction of our training set. Figure 5.32 describe the model output and FTSE 100 variety of texting data. Similar like training model, the blue line describes the FTSE data and the red line is the wavelet model output. The result of the wavelet model performance is: the root mean square error is 11.7407 and the mean absolute error is 9.5484 for training model.

1. Twitter Sentiment & Emotion and FTSE model

In this part, we have combined the wavelet decomposed Twitter sentiment and Twitter emotion together to model FTSE100 close price. Results are shown in figures below:



Figure 5. 33 Simulation results of daily FTSE Twitter sentiment & emotion model



Figure 5. 34 Simulation results of daily FTSE Twitter sentiment & emotion model validation

The regressor terms are chosen by OLS + ERR method. Set the threshold to 0.001, the selection results of Twitter sentiment & emotion model are: A3\_FTSE, D3\_Trust, D1\_FTSE, D2\_FTSE, D2\_Surprise, D2\_Disgust, D1\_Anger, A3\_Positive. The simulation results of training and testing are shown in figures above. As we can see from Figure 5.33, the blue line describes the FTSE 100 close price change and the red line is the one step ahead prediction of our training set. Figure 5.34 describe the model output and FTSE 100 variety of texting data. Similar like training model, the blue line describes the FTSE data and the red line is the wavelet model output. The result of the wavelet model performance is: the root mean square error is 17.4576 and the mean absolute error is 13.5 for training model.

1. Results Analysis and Summary

In this part, researchers have applied FTSE Twitter sentiment index and FTSE Twitter emotion index and these data will partly reflect the public’s altitudes towards UK stock market. Although Twitter sentiment/emotion data only represent part of the public opinion, the Twitter data have shown strong predictive power to model the UK stock market variety. Furthermore, the predictive performance of the Twitter emotion index is even better than the Twitter sentiment index. The wavelet based NARX model performance will be prepared with NARX model in Table below.

Table 5. 11 The performance of Wavelet NARX and NARX about Twitter FTSE system

|  |  |  |
| --- | --- | --- |
|  | MAE | RMSE |
| Wavelet Sentiment | 11.1159 | 14.0519 |
| Wavelet Emotion | 11.7407 | 9.5484 |
| Sentiment | 22.9090 | 32.8863 |
| Emotion | 17.9364 | 24.4050 |

In this part, we compared the performance of Wavelet based NARX model and NARX model on Twitter Sentiment/Emotion FTSE system. The results show that, compared with NARX, wavelet can significantly improve the model performance of FTSE Twitter system. Which prove that Wavelet pre-process is an important process in modelling severely nonlinear and un-stationary stock market system.

# Chapter 6. Conclusion

In this PhD project, researchers mainly focus on using Twitter data and system identification techniques to model and predict the real world non-linear and non-stationary process, such as stock market system. In the process of modelling and predicting these economic systems, researchers find that microblogging on the Internet contains a lot of sentiment and emotion information. The tweets information includes economics such as stock market and political issue such as presidential election. Twitter, as one of the most popular social network services, can provide opinion-rich tweets information for our experiment. Based on behaviour economics, stock market price change is usually driven by the sentiments of stock investors. Therefore, researchers will use Twitter sentiment data to model the real world non-linear and un-stationary stock market system. In general, this project mainly investigates system identification methods, data mining and text mining, lexicon based method, wavelet analysis, complex network analysis and machine learning algorithms in our Twitter stock market systems.

The extraction of Twitter data is difficult and expensive, three methods Googlespread Sheets, Webharvey and Twitter API in R is used to extract tweets from Twitter. The experiment results show that, considering about the data integrity and data diversity, we have developed a program in R based on Twitter API, it performs better than Googlespread Sheet and Webharvey. Furthermore, R can store the extracted tweets in either excel format or *.Rdata*, and it is convenience for our future experiments. In addition, Twitter API in R can also extract the tweets data based on the geography location, and the geography Twitter data can help us to investigate Twitter opinion about Brexit 2016. The geographic information has made it possible for us to understand Twitter public sentiment about Brexit in a comprehensive perspective. Lastly, the Twitter API in R can extract the update/real-time Twitter data, this property will make our experiment more efficient. Although R cannot perform daily extraction tasks automatically and researchers need to extract Twitter data day by day, Twitter API in R has successfully extracted 3 million of US presidential election Twitter data; more than twenty thousand of Brexit Twitter data, more than 90000 FTSE 100 Twitter data.

There is an important task of this project is to mine the sentiment/emotion index from Twitter. The tweets are related to US presidential election, Brexit 2016 and FTSE 100.

We have made a novel application of NRC Lexicon on the semantic analysis of US presidential election Twitter data, Brexit Twitter data and FTSE 100 Twitter. We have obtained valuable public opinion information for presidential election and UK referendum. The sentiment and emotion index distribution of the two presidential candidates before the election which is proven to be which is close to real world situation. For example, the daily surprise emotion index of Donald Trump is significantly higher than Hillary Clinton on every single day before the election date. In general, the Twitter opinion results show Trump has the higher emotion index on Twitter than Hillary which means that the Twitter related to Trump gives us more emotions words. By summarize the sentiment and emotion index of these two presidential candidates, results show that Trump is more competitiveness on Twitter than Hillary. Although our Twitter model results show that the UK Brexit referendum 2016 is more people support stay in European Union, however, the real referendum results is opposite. The reasons are: 1. There is not enough Twitter samples for our experiments; 2. The Brexit Twitter has not been extracted day by day, which make us cannot see the changes in public opinions; 3. Not everyone use Twitter to express their opinions. It is believed that by deep mining these Twitter data, we can obtain more information on public opinion. With the help of NRC lexicon, we also get Twitter sentiment indexes and Twitter emotion indexes about FTSE100. For the future research, these opinion-rich datasets can help us to modelling economic problem based on nonlinear models and complex network theory.

We have also investigated the development and defects of current sentiment analysis methodologies; Summarizes the research status of current text classification method; Twitter text data pre-processing technology; applied proposed improved lexicon based method on Twitter economic data and Twitter political data; proposed a novel feature selection method on KNN and Naïve Bayes. Some economic and political topics’ sentiment distribution on Twitter has been visualized, the results have made it possible for researchers to understand the public opinion of these topics. This project has also developed a text classification system that include training, classification and evaluation processes. This method is able to complete the entire process of Twitter sentiment analysis. A combination of NRC feature selection methods and KNN, Naïve Bayes classifier is developed. The experiment results show that the performance of classification results shows that NRC KNN outperform than NRC Naïve Bayes.

This project has done numerous studies on background research, theoretical research, system design, modelling and argument process about how to model the stock market change based on, crude oil price and Twitter public sentiment index. Researchers implemented linear and nonlinear Wavelet models and the sentiment time series to model FTSE100 system. The main results show that: 1. Compared with other system identification methods, compared the model that without wavelet, Wavelet NARX model can significantly improve the prediction power for stock market system. 2. Short term prediction of the oil price perform better than long term model. 3. Twitter sentiment and Twitter emotion can help us to predict the FTSE100 change. A novel methodology that implement Twitter sentiment data to model non-linear and non-stationary FTSE100 system is developed and this algorithm can also be used in other economic system or political election system.

With the development of the social networks service, various types of Tweets data have attracted the attention of researchers and the corresponding research work also has potential in economic system, political issue and public opinion monitor. This project has preliminarily explored the influence online data in modelling the real world political and economic system, considering about development of big data and the applicability of this algorithm, this project still has potential in future research:

* In stock market research, a software platform that is able to extract and mining online text sentiment data automatically is a further research direction. This software will provide important reference for stock market research. With the design and development of this software, a platform that based on Twitter public opinion and stock market variety can be applied into practice.
* In forecasting the stock market price volatility, Economic decision analysis, Risk assessment and management method can be introduced to help the government to supervise and control the financial market.
* The Twitter sentiment analysis can be extended to different fields, which contains several commercial values. In E-commerce, this method could help manufactures and companies to understand the online public sentiment information about the commodities and products. In public opinion control, this algorithm can help the government to understand and control the public opinion. This can help the government to prevent malicious rumours and understand the public opinion in some major social issues.

# References

Adamo, F., Andria, G., Attivissimo, F., Lanzolla, A.M.L. and Spadavecchia, M., 2013. A comparative study on mother wavelet selection in ultrasound image denoising. *Measurement*, *46*(8), pp.2447-2456.

Adamowski, J., & Sun, K. (2010). Development of a coupled wavelet transform and neural network method for flow forecasting of non-perennial rivers in semi-arid watersheds. *Journal of Hydrology*, *390*(1), 85-91.

Aggarwal, C.C. and Zhai, C. eds., 2012. *Mining text data*. Springer Science & Business Media.

Ahadi, M. and Bakhtiar, M.S., 2010. Leak detection in water-filled plastic pipes through the application of tuned wavelet transforms to acoustic emission signals. *Applied Acoustics*, *71*(7), pp.634-639.

Ahire, S. (2014). A Survey of Sentiment Lexicons.

Ahuja, N., Lertrattanapanich, S. and Bose, N.K., 2005. Properties determining choice of mother wavelet. *IEE Proceedings-Vision, Image and Signal Processing*, *152*(5), pp.659-664.

Al-Qazzaz, N.K., Hamid Bin Mohd Ali, S., Ahmad, S.A., Islam, M.S. and Escudero, J., 2015. Selection of mother wavelet functions for multi-channel eeg signal analysis during a working memory task. *Sensors*, *15*(11), pp.29015-29035.

Al Wadia, M.T.I.S. and Tahir Ismail, M., 2011. Selecting wavelet transforms model in forecasting financial time series data based on ARIMA model. *Applied Mathematical Sciences*, *5*(7), pp.315-326.

Alquist, R., Kilian, L., & Vigfusson, R. (2011). Forecasting the price of oil. *Available at SSRN 1911194*.

Apte, C., Damerau, F. and Weiss, S., 1998. *Text mining with decision rules and decision trees*. IBM Thomas J. Watson Research Division.

Arafat, S.M., 2003. *Uncertainty modeling for classification and analysis of medical signals* (Doctoral dissertation, University of Missouri-Columbia).

Billings, S. A. (2013). *Nonlinear system identification: NARMAX methods in the time, frequency, and spatio-temporal domains*. John Wiley & Sons.

Billings, S. A., & Wei, H. L. (2005). The wavelet-NARMAX representation: A hybrid model structure combining polynomial models with multiresolution wavelet decompositions. *International Journal of Systems Science*, *36*(3), 137-152.

Bollen, J., Mao, H., & Pepe, A. (2011, July). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *ICWSM*.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1-8.

Bordino, I., Battiston, S., Caldarelli, G., Cristelli, M., Ukkonen, A., & Weber, I. (2012). Web search queries can predict stock market volumes. PloS one, 7(7), e40014.

Bijalwan, V., Kumar, V., Kumari, P. and Pascual, J., 2014. KNN based machine learning approach for text and document mining. *International Journal of Database Theory and Application*, *7*(1), pp.61-70.

Bird, D., Ling, M. and Haynes, K., 2012. Flooding Facebook-the use of social media during the Queensland and Victorian floods. Australian Journal of Emergency Management, The, 27(1), p.27.

Bonsignore, E.M., Dunne, C., Rotman, D., Smith, M., Capone, T., Hansen, D.L. and Shneiderman, B., 2009, August. First steps to NetViz Nirvana: evaluating social network analysis with NodeXL. In *Computational Science and Engineering, 2009. CSE'09. International Conference on* (Vol. 4, pp. 332-339). IEEE.

Brennan, S., Sadilek, A., & Kautz, H. (2013, August). Towards understanding global spread of disease from everyday interpersonal interactions. In Proceedings of the Twenty-Third international joint conference on Artificial Intelligence (pp. 2783-2789). AAAI Press.

Campbell, J. Y., & Thompson, S. B. (2005). *Predicting the Equity Premium out of sample: Can anything beat the historical average?* (No. w11468). National Bureau of Economic Research.

Cao, Q., Leggio, K. B., & Schniederjans, M. J. (2005). A comparison between Fama and French's model and artificial neural networks in predicting the Chinese stock market. *Computers & Operations Research*, *32*(10), 2499-2512.

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, P. K. (2010). Measuring user influence in twitter: The million follower fallacy. *Icwsm*, *10*(10-17), 30.

Chen, R. and Lazer, M., 2013. Sentiment analysis of twitter feeds for the prediction of stock market movement. *stanford. edu. Retrieved January*, *25*, p.2013.

Chiras, N., Evans, C., & Rees, D. (2001). Nonlinear gas turbine modeling using NARMAX structures. *Instrumentation and Measurement, IEEE Transactions on*, *50*(4), 893-898.

Chen, S., & Billings, S. A. (1989). Representations of non-linear systems: the NARMAX model. *International Journal of Control*, *49*(3), 1013-1032.

Coca, D., & Billings, S. A. (2001). Non-linear system identification using wavelet multiresolution models. *International Journal of Control*, *74*(18), 1718-1736.

Cohen, W. W., & Singer, Y. (1999). Context-sensitive learning methods for text categorization. *ACM Transactions on Information Systems (TOIS)*, *17*(2), 141-173.

Cooley, R., Mobasher, B. and Srivastava, J., 1997, November. Web mining: Information and pattern discovery on the world wide web. In *Tools with Artificial Intelligence, 1997. Proceedings., Ninth IEEE International Conference on* (pp. 558-567). IEEE.

Dang, Y., Zhang, Y. and Chen, H., 2010. A lexicon-enhanced method for sentiment classification: An experiment on online product reviews. *IEEE Intelligent Systems*, *25*(4), pp.46-53.

Deng, L., & Tan, Y. (2009). Modeling hysteresis in piezoelectric actuators using NARMAX models. *Sensors and Actuators A: Physical*, *149*(1), 106-112.

Dilrukshi, I., De Zoysa, K. and Caldera, A., 2013, April. Twitter news classification using SVM. In *Computer Science & Education (ICCSE), 2013 8th International Conference on* (pp. 287-291). IEEE.

Dolan, R.J., 2002. Emotion, cognition, and behavior. *science*, *298*(5596), pp.1191-1194.

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

Eichler, M. (2012). Causal inference in time series analysis. *Causality: Statistical perspectives and applications*, 327-354.

Eisenstein, J. (2017). Unsupervised Learning for Lexicon-Based Classification. In *AAAI* (pp. 3188-3194).

Enke, D., & Thawornwong, S. (2005). The use of data mining and neural networks for forecasting stock market returns. Expert Systems with applications, 29(4), 927-940.

Fama, E.F., 1965. The behavior of stock-market prices. *The journal of Business*, *38*(1), pp.34-105.

Flanders, M., 2002. Choosing a wavelet for single-trial EMG. *Journal of neuroscience methods*, *116*(2), pp.165-177.

Ferreira, M. A., & Santa-Clara, P. (2011). Forecasting stock market returns: The sum of the parts is more than the whole. *Journal of Financial Economics*,*100*(3), 514-537.

Fung, E. H., Wong, Y. K., Ho, H. F., & Mignolet, M. P. (2003). Modelling and prediction of machining errors using ARMAX and NARMAX structures. *Applied Mathematical Modelling*, *27*(8), 611-627.

Fu, S., Muralikrishnan, B. and Raja, J., 2003. Engineering surface analysis with different wavelet bases. *TRANSACTIONS-AMERICAN SOCIETY OF MECHANICAL ENGINEERS JOURNAL OF MANUFACTURING SCIENCE AND ENGINEERING*, *125*(4), pp.844-852.

García, A., Gaines, S., & Linaza, M. T. (2012). A lexicon based sentiment analysis retrieval system for tourism domain. *Expert Syst Appl Int J*, *39*(10), 9166-9180.

Gayathri, K. and Marimuthu, A., 2013, January. Text document pre-processing with the KNN for classification using the SVM. In *Intelligent Systems and Control (ISCO), 2013 7th International Conference on* (pp. 453-457). IEEE.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2008). Detecting influenza epidemics using search engine query data. Nature, 457(7232), 1012-1014.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1-12.

Gräbner, D., Zanker, M., Fliedl, G., & Fuchs, M. (2012). *Classification of customer reviews based on sentiment analysis* (pp. 460-470). na.

Gupta, V. and Lehal, G.S., 2009. A survey of text mining techniques and applications. *Journal of emerging technologies in web intelligence*, *1*(1), pp.60-76.

Hamid, E.Y., Mardiana, R. and Kawasaki, Z.I., 2001, July. Wavelet-based compression of power disturbances using the minimum description length criterion. In *Power Engineering Society Summer Meeting, 2001* (Vol. 3, pp. 1772-1777). IEEE.

Hemalatha, I., Varma, G.S. and Govardhan, A., 2012. Preprocessing the informal text for efficient sentiment analysis. *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, *1*(2), pp.58-61.

Honey, C. and Herring, S.C., 2009, January. Beyond microblogging: Conversation and collaboration via Twitter. In *System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on* (pp. 1-10). IEEE.

Hsieh, T. J., Hsiao, H. F., & Yeh, W. C. (2011). Forecasting stock markets using wavelet transforms and recurrent neural networks: An integrated system based on artificial bee colony algorithm. *Applied soft computing*, *11*(2), 2510-2525.

Hu, C.A., & Zhu, LJ. (2010). The evaluation and analysis of complex network software. Digital library ISTIC, (5).

Hu, S., & Liang, H. (2012). Causality analysis of neural connectivity: New tool and limitations of spectral granger causality. *Neurocomputing*, *76*(1), 44-47.

Hu, X., Tang, J., Gao, H., & Liu, H. (2013, May). Unsupervised sentiment analysis with emotional signals. In Proceedings of the 22nd international conference on World Wide Web (pp. 607-618). International World Wide Web Conferences Steering Committee.

Hsieh, T.J., Hsiao, H.F. and Yeh, W.C., 2011. Forecasting stock markets using wavelet transforms and recurrent neural networks: An integrated system based on artificial bee colony algorithm. *Applied soft computing*, *11*(2), pp.2510-2525.

Jain, A., & Kumar, A. M. (2007). Hybrid neural network models for hydrologic time series forecasting. *Applied Soft Computing*, *7*(2), 585-592.

Jansen, B.J., Zhang, M., Sobel, K. and Chowdury, A., 2009. Twitter power: Tweets as electronic word of mouth. *Journal of the Association for Information Science and Technology*, *60*(11), pp.2169-2188.

Jiang, L., Yu, M., Zhou, M., Liu, X. and Zhao, T., 2011, June. Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (pp. 151-160). Association for Computational Linguistics.

Jurek, A., Mulvenna, M. D., & Bi, Y. (2015). Improved lexicon-based sentiment analysis for social media analytics. *Security Informatics*, *4*(1), 9.

Kahneman, D. and Tversky, A., 2013. Prospect theory: An analysis of decision under risk. In *HANDBOOK OF THE FUNDAMENTALS OF FINANCIAL DECISION MAKING: Part I* (pp. 99-127).

Kankar, P.K., Sharma, S.C. and Harsha, S.P., 2011. Fault diagnosis of ball bearings using continuous wavelet transform. *Applied Soft Computing*, *11*(2), pp.2300-2312.

Kantardzic, M., 2011. *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons.

Kao, L.J., Chiu, C.C., Lu, C.J. and Chang, C.H., 2013. A hybrid approach by integrating wavelet-based feature extraction with MARS and SVR for stock index forecasting. *Decision Support Systems*, *54*(3), pp.1228-1244.

Katariya, M.N.P., Chaudhari, M.S., Subhani, B., Laxminarayana, G., Matey, K., Nikose, M.A., Tinkhede, S.A. and Deshpande, S.P., 2015. Text preprocessing for text mining using side information. *International Journal of Computer Science and Mobile Applications*, *3*(1), pp.01-05.

Khan, A., Baharudin, B., Lee, L.H. and Khan, K., 2010. A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology*, *1*(1), pp.4-20.

Kim, H. J., & Shin, K. S. (2007). A hybrid approach based on neural networks and genetic algorithms for detecting temporal patterns in stock markets. Applied Soft Computing, 7(2), 569-576.

Kosala, R. and Blockeel, H., 2000. Web mining research: A survey. *ACM Sigkdd Explorations Newsletter*, *2*(1), pp.1-15.

Kouloumpis, E., Wilson, T. and Moore, J.D., 2011. Twitter sentiment analysis: The good the bad and the omg!. *Icwsm*, *11*(538-541), p.164.

Kuo, C. C., Gan, T. Y., & Yu, P. S. (2010). Seasonal streamflow prediction by a combined climate-hydrologic system for river basins of Taiwan. *Journal of hydrology*, *387*(3), 292-303.

Larsson, A.O. and Moe, H., 2012. Studying political microblogging: Twitter users in the 2010 Swedish election campaign. *New Media & Society*, *14*(5), pp.729-747.

Lenhart, A., Purcell, K., Smith, A., & Zickuhr, K. (2010). Social media and young adults. Pew Internet & American Life Project, 3.

Li, N. and Wu, D.D., 2010. Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision support systems*, *48*(2), pp.354-368

Li, W., 2009, February. Research on extraction of partial discharge signals based on wavelet analysis. In *Electronic Computer Technology, 2009 International Conference on* (pp. 545-548). IEEE.

Li, Y., Wei, H. L., Billings, S. A., & Liao, X. F. (2012). Time-varying linear and nonlinear parametric model for Granger causality analysis. *Physical Review E*,*85*(4), 041906.

Li Ying, Zhang Xiaohui, Wang Huayong and Chang Guiran, 2004. A Chinese Text Classification Method based on Vector Aggregation. *Small Microcomputer System*, *25*(6), pp.993-996.

Liu, B., Hu, M. and Cheng, J., 2005, May. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web* (pp. 342-351). ACM.

Liu, B. (2010). Sentiment Analysis and Subjectivity. *Handbook of natural language processing*, *2*, 627-666.

Liu, B., 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, *5*(1), pp.1-167.

Luhn, H.P., 1957. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of research and development*, *1*(4), pp.309-317.

Ferreira, M.A. and Santa-Clara, P., 2011. Forecasting stock market returns: The sum of the parts is more than the whole. Journal of Financial Economics, 100(3), pp.514-537.

Maheswaran, R., & Khosa, R. (2014). A wavelet-based second order nonlinear model for forecasting monthly rainfall. *Water Resources Management*, *28*(15), 5411-5431.

Makwana, J. J., & Tiwari, M. K. (2014). Intermittent Streamflow Forecasting and Extreme Event Modelling using Wavelet based Artificial Neural Networks. *Water Resources Management*, *28*(13), 4857-4873.

Mallat, S. (2008). A wavelet tour of signal processing: the sparse way. Academic press.

Mangold, W. G., & Faulds, D. J. (2009). Social media: The new hybrid element of the promotion mix. Business horizons, 52(4), 357-365.

Mao, H., Counts, S., & Bollen, J. (2011). Predicting financial markets: Comparing survey, news, twitter and search engine data. arXiv preprint arXiv:1112.1051.

Mao, Y., Wei, W., Wang, B. and Liu, B., 2012, August. Correlating S&P 500 stocks with Twitter data. In *Proceedings of the first ACM international workshop on hot topics on interdisciplinary social networks research* (pp. 69-72). ACM.

Marg, E., 1995. DESCARTES'ERROR: Emotion, Reason, and the Human Brain. *Optometry & Vision Science*, *72*(11), pp.847-848.

Miao, K., Chen, F. and Zhao, Z.G., 2007. Stock price forecast based on bacterial colony RBF neural network [j]. *Journal of Qingdao University (Natural Science Edition)*, *20*(2), pp.50-54.

Mojsilovic, A., Popovic, M.V. and Rackov, D.M., 2000. On the selection of an optimal wavelet basis for texture characterization. *IEEE Transactions on Image Processing*, *9*(12), pp.2043-2050.

Mohammad, S. M., & Turney, P. D. (2013). *Nrc emotion lexicon*. NRC Technical Report.

Moraes, R., Valiati, J.F. and Neto, W.P.G., 2013. Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Systems with Applications*, *40*(2), pp.621-633.

Murray, G., Joty, S. R., Carenini, G., & Ng, R. T. (2008). The University of British Columbia at TAC 2008. In *TAC*.

Narayanan, V., Arora, I. and Bhatia, A., 2013, October. Fast and accurate sentiment classification using an enhanced Naive Bayes model. In *International Conference on Intelligent Data Engineering and Automated Learning* (pp. 194-201). Springer Berlin Heidelberg.

Nasukawa, T. and Yi, J., 2003, October. Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on Knowledge capture* (pp. 70-77). ACM.

Ngui, W.K., Leong, M.S., Hee, L.M. and Abdelrhman, A.M., 2013. Wavelet analysis: mother wavelet selection methods. In *Applied mechanics and materials* (Vol. 393, pp. 953-958). Trans Tech Publications.

Nofer, M. and Hinz, O., 2015. Using twitter to predict the stock market. *Business & Information Systems Engineering*, *57*(4), pp.229-242.

Nofsinger, J. R. (2005). Social mood and financial economics. The Journal of Behavioral Finance, 6(3), 144-160.

Ortigosa, A., Martín, J.M. and Carro, R.M., 2014. Sentiment analysis in Facebook and its application to e-learning. *Computers in Human Behavior*, *31*, pp.527-541.

Pak, A., & Paroubek, P. (2010, May). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In LREC.

Petz, G., Karpowicz, M., Fürschuß, H., Auinger, A., Winkler, S.M., Schaller, S. and Holzinger, A., 2012, December. On text preprocessing for opinion mining outside of laboratory environments. In *International Conference on Active Media Technology* (pp. 618-629). Springer Berlin Heidelberg.

Phinyomark, A., Limsakul, C. and Phukpattaranont, P., 2009. A novel feature extraction for robust EMG pattern recognition. arXiv preprint arXiv:0912.3973.

Pol, K., Patil, N., Patankar, S. and Das, C., 2008, July. A Survey on Web Content Mining and extraction of Structured and Semistructured data. In *Emerging Trends in Engineering and Technology, 2008. ICETET'08. First International Conference on* (pp. 543-546). IEEE.

Prechter Jr, R. R. (2002). The Wave Principle of Human Social Behavior and the: New Science of Socionomics (Vol. 1). New Classics Library.

Qian, B. and Rasheed, K., 2007. Stock market prediction with multiple classifiers. *Applied Intelligence*, *26*(1), pp.25-33.

Rahrooh, A., & Shepard, S. (2009). Identification of nonlinear systems using NARMAX model. *Nonlinear Analysis: Theory, Methods & Applications*, *71*(12), e1198-e1202.

Rua, A. and Nunes, L.C., 2009. International comovement of stock market returns: A wavelet analysis. *Journal of Empirical Finance*, *16*(4), pp.632-639.

Sadilek, A., Kautz, H. A., & Silenzio, V. (2012, June). Modeling Spread of Disease from Social Interactions. In ICWSM.

Safavian, L.S., Kinsner, W. and Turanli, H., 2005, May. A quantitative comparison of different mother wavelets for characterizing transients in power systems. In *Electrical and Computer Engineering, 2005. Canadian Conference on* (pp. 1461-1464). IEEE.

Saito, N., 1994, March. Simultaneous noise suppression and signal compression using a library of orthonormal bases and the minimum-description-length criterion. In *SPIE's International Symposium on Optical Engineering and Photonics in Aerospace Sensing* (pp. 224-235). International Society for Optics and Photonics.

Schmid, H., 1995. Treetagger| a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, *43*, p.28.

Schoen, H., Gayo-Avello, D., Takis Metaxas, P., Mustafaraj, E., Strohmaier, M. and Gloor, P., 2013. The power of prediction with social media. Internet Research, 23(5), pp.528-543.

Shik Lee, H., 2004. International transmission of stock market movements: a wavelet analysis. *Applied Economics Letters*, *11*(3), pp.197-201.

Singh, B.N. and Tiwari, A.K., 2006. Optimal selection of wavelet basis function applied to ECG signal denoising. *Digital signal processing*, *16*(3), pp.275-287.

Singh, V.K., Piryani, R., Uddin, A. and Waila, P., 2013, March. Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification. In *Automation, computing, communication, control and compressed sensing (iMac4s), 2013 international multi-conference on* (pp. 712-717). IEEE.

Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H. and Deng, X., 2013. Exploiting Topic based Twitter Sentiment for Stock Prediction. *ACL (2)*, *2013*, pp.24-29.

Sparck Jones, K., 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, *28*(1), pp.11-21.

Taboada, M., Brooke, J., & Stede, M. (2009, September). Genre-based paragraph classification for sentiment analysis. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 62-70). Association for Computational Linguistics.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, *37*(2), 267-307.

Tan, T. Z., Quek, C., & Ng, G. S. (2005, September). Brain-inspired genetic complementary learning for stock market prediction. In *Evolutionary Computation, 2005. The 2005 IEEE Congress on* (Vol. 3, pp. 2653-2660). IEEE.

Tang, B., Liu, W. and Song, T., 2010. Wind turbine fault diagnosis based on Morlet wavelet transformation and Wigner-Ville distribution. *Renewable Energy*, *35*(12), pp.2862-2866.

Tang, D., Qin, B. and Liu, T., 2015, September. Document Modeling with Gated Recurrent Neural Network for Sentiment Classification. In *EMNLP* (pp. 1422-1432).

Ting, S.L., Ip, W.H. and Tsang, A.H., 2011. Is Naive Bayes a good classifier for document classification. *International Journal of Software Engineering and Its Applications*, *5*(3), pp.37-46.

Trstenjak, B., Mikac, S. and Donko, D., 2014. KNN with TF-IDF based Framework for Text Categorization. *Procedia Engineering*, *69*, pp.1356-1364.

Uysal, A.K. and Gunal, S., 2014. The impact of preprocessing on text classification. *Information Processing & Management*, *50*(1), pp.104-112.

Venezia, I., Nashikkar, A., & Shapira, Z. (2011). Firm specific and macro herding by professional and amateur investors and their effects on market volatility. Journal of Banking & Finance, 35(7), 1599-1609.

Wang, H., Can, D., Kazemzadeh, A., Bar, F. and Narayanan, S., 2012, July. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In Proceedings of the ACL 2012 System Demonstrations (pp. 115-120). Association for Computational Linguistics.

Wang, S.Y., Liu, X., Yianni, J., Aziz, T.Z. and Stein, J.F., 2004. Extracting burst and tonic components from surface electromyograms in dystonia using adaptive wavelet shrinkage. *Journal of neuroscience methods*, *139*(2), pp.177-184.

Wang, X., Wei, F., Liu, X., Zhou, M. and Zhang, M., 2011, October. Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In *Proceedings of the 20th ACM international conference on Information and knowledge management* (pp. 1031-1040). ACM.

Wei, H. L., & Billings, S. A. (2002). Identification of time-varying systems using multiresolution wavelet models. *International Journal of Systems Science*,*33*(15), 1217-1228.

Wei, H. L., & Billings, S. A. (2004). A unified wavelet-based modelling framework for non-linear system identification: the WANARX model structure. *International Journal of Control*, *77*(4), 351-366.

Wei, H. L., Billings, S. A., & Balikhin, M. (2004). Prediction of the Dst index using multiresolution wavelet models. *Journal of Geophysical Research: Space Physics (1978–2012)*, *109*(A7).

Wei, H. L., & Billings, S. A. (2006). Long term prediction of non-linear time series using multiresolution wavelet models. *International Journal of Control*,*79*(06), 569-580.

Wei, H. L., & Billings, S. A. (2009). Power-law behaviour evaluation from foreign exchange market data using a wavelet transform method. *Physics Letters A*, *373*(37), 3324-3329.

Wei, H. L., Billings, S. A., & Liu, J. J. (2010). Time-varying parametric modelling and time-dependent spectral characterisation with applications to EEG signals using multiwavelets. *International Journal of Modelling, Identification and Control*, *9*(3), 215-224.

Weiss, S.M., Indurkhya, N., Zhang, T. and Damerau, F., 2010. *Text mining: predictive methods for analyzing unstructured information*. Springer Science & Business Media.

Xianghua, F., Guo, L., Yanyan, G. and Zhiqiang, W., 2013. Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon. *Knowledge-Based Systems*, *37*, pp.186-195.

Yan, R. and Gao, R.X., 2009. Base wavelet selection for bearing vibration signal analysis. *International Journal of Wavelets, Multiresolution and Information Processing*, *7*(04), pp.411-426.

Yang, Z., Guo, J., Cai, K., Tang, J., Li, J., Zhang, L. and Su, Z., 2010, October. Understanding retweeting behaviors in social networks. In Proceedings of the 19th ACM international conference on Information and knowledge management (pp. 1633-1636). ACM.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A.J. and Hovy, E.H., 2016. Hierarchical Attention Networks for Document Classification. In *HLT-NAACL* (pp. 1480-1489).

Yong, Z., Youwen, L. and Shixiong, X., 2009. An improved KNN text classification algorithm based on clustering. *Journal of computers*, *4*(3), pp.230-237.

Yue Yunfei, Wang Wei, Liu Dayou and Shao Liangshan, 2012. Feature Selection Method Based on Variance CHI. *Computer Application Research*, *29*(4), pp.1304-1306.

Zhang, L., Bao, P. and Wu, X., 2005. Multiscale LMMSE-based image denoising with optimal wavelet selection. *IEEE Transactions on circuits and systems for video technology*, *15*(4), pp.469-481.

Zhang, Y., & Wu, L. (2009). Stock market prediction of S&P 500 via combination of improved BCO approach and BP neural network. Expert systems with applications, 36(5), 8849-8854.